Originalveröffentlichung in: Deppermann, Arnulf/Fandrych, Christian/Kupietz, Marc/Schmidt, Thomas (Hrsg.): Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial. – Berlin/Boston: de Gruyter, 2023. S. 199-223. (Jahrbuch / Leibniz-Institut für Deutsche Sprache 2022)
DOI: *https://doi.org/10.1515/9783111085708*

Volker Emmrich/Mathilde Hennig (Gießen)

GiesKaNe: Korpusaufbau zwischen Standard und Innovation

Abstract: Der vorliegende Beitrag erörtert am Beispiel des aktuell im Aufbau befindlichen Korpus GiesKaNe (= Gie[ßen]Ka[ssel]Ne[uhochdeutsch]) grundlegende Fragen nach dem Verhältnis von Standard und Innovation bei der Erweiterung der Korpuslandschaft durch neue Korpora. Bei jedem neu zu erstellenden Korpus stellt sich die Frage, inwieweit man den bereits etablierten Standards folgt, oder ob es legitim oder vielleicht sogar notwendig ist, neue Modelle der Annotation linguistischer Kategorien zu entwickeln. In diesem Sinne bespricht der Beitrag die Grenzen einer reinen Modellübernahme mit Bezug auf das POS-Tagging in anderen historischen Referenzkorpora und mit Bezug auf TIGER als Baumbank für das Gegenwartsdeutsche. Um trotz der Arbeit mit einer innovativen Alternative dem Prinzip der Interoperabilität gerecht zu werden, wird im Beitrag die Arbeit mit maschinellem Lernen ins Spiel gebracht. Dieses ermöglicht es, aus den vorhandenen Textoberflächenmerkmalen und den vorliegenden Annotationen auch alternative Annotationsmodelle abzuleiten und mittels einer Mehrebenenannotation anzubieten, sodass ein Korpus den Anforderungen an interoperable Nutzbarkeit und wissenschaftlichen Erkenntnisfortschritt gleichermaßen gerecht werden kann.

1 Einleitung

Der vorliegende Beitrag erörtert am Beispiel des aktuell im Aufbau befindlichen Korpus GiesKaNe (= Gie[ßen]Ka[ssel]Ne[uhochdeutsch]) grundlegende Fragen nach dem Verhältnis von Standard und Innovation bei der Erweiterung der Korpuslandschaft durch neue Korpora. Es besteht in der (Korpus-)Linguistik aktuell ein breiter Konsens in Bezug auf die Notwendigkeit einer Orientierung an Standards: Man denke nur an die im Grunde flächendeckende Nutzung des STTS zur Wortartannotation oder das breite Bekenntnis zu TEI. Die Standardorientierung bietet zweifelsohne klare Vorteile sowohl für die Korpuserstellung als auch die Korpusnutzung: In der Korpuserstellung muss das Rad nicht jedes Mal neu erfunden werden, der Korpusersteller kann auf Bestehendes zurückgreifen und sich auf diese Weise voll und ganz auf sein Forschungsinteresse konzentrieren. Korpusnutzer/-innen können auf ihren Vorkenntnissen zu Korpora aufbauen und müs-

https://doi.org/10.1515/9783111085708-010

sen sich nicht bei jeder Nutzung eines neuen Korpus erneut in Tagsets und Annotationsmodelle einarbeiten. Schließlich bieten Standards die Grundlage für die interoperable Nutzung von Korpora, also die Bearbeitung einer Fragestellung mit Hilfe mehrerer Korpora – was natürlich voraussetzt, dass in diesen Korpora die gleichen Analysekategorien durch Annotation zugänglich gemacht wurden.

Das aktuell – soweit wir es überblicken – kaum diskutierte und hinterfragte Modell der Standardorientierung steht jedoch in einem grundlegenden Konflikt mit dem für wissenschaftlichen Fortschritt zentralen Prinzip der Innovation. Standardorientierung in der Korpuserstellung bedeutet im Grunde genommen, dass ein zu einem bestimmten Zeitpunkt aus bestimmten Gründen festgelegtes Modell multipliziert wird. Nach Standards erschlossene Korpora erhöhen die Datenmenge für die Analyse von Sprachdaten mit diesen Modellen. Auch wenn diese Vorgehensweise durchaus für einen Erkenntnisfortschritt sorgen kann – etwa in dem Sinne, dass man zu Aussagen der Verwendung eines sprachlichen Phänomens unter verschiedenen pragmatischen, historischen und medialen Bedingungen gelangt – sind Standards aus der Perspektive des wissenschaftlichen Erkenntnisinteresses dann problematisch, wenn sie als unabänderlicher Endpunkt einer Entwicklung begriffen werden. Hinzu kommt in der Korpuslinguistik auch, dass die Entwicklung von Standards auf der Basis der zum jeweiligen Zeitpunkt vorliegenden computermethodischen Möglichkeiten erfolgt, die selbstverständlich auch einer Entwicklung unterliegen. Dieser eher technische Aspekt des Verhältnisses von Standard und Innovation steht allerdings nicht im Fokus unseres Beitrags. Uns interessiert vielmehr das folgende grundlegende Dilemma: Während Wissenschaft das Bestehende diskutiert und erweitert, muss ein Korpus zunächst das bestehende Wissen in Form von Annotationsmodellen aufgreifen. Dabei soll gerade das Korpus als Datengrundlage für die Generierung neuen Wissens dienen. Es darf also nicht nur konservativ auf die Zementierung von Bestehendem ausgerichtet sein, sondern es sollte auch einen Möglichkeitsraum für wissenschaftliche Innovationen bieten.

Das im Aufbau befindliche Korpus GiesKaNe setzt in diesem Sinne auf Innovation. Das Ziel des vorliegenden Beitrags besteht darin, am Beispiel von GiesKaNe zu zeigen, wie – insbesondere auf der Basis der Möglichkeiten des maschinellen Lernens – gerade auch innovative Ansätze für die Rekonstruktion von Standards genutzt werden können, damit ein Korpus gleichermaßen den skizzierten Anforderungen an eine interoperable Nutzung und an den Erkenntnisfortschritt gerecht werden kann.

2 Anforderungen an ein (Referenz-)Korpus

Das Korpus GiesKaNe befindet sich seit 2016 im Rahmen des von der DFG geförderten, von Vilmos Ágel (Universität Kassel) und Mathilde Hennig (JLU Gießen) geleiteten Langfristvorhabens „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch“ im Aufbau. Für eine Diskussion der Anforderungen, auf die der Korpusaufbau von GiesKaNe zu reagieren hat, beginnen wir mit einer Auseinandersetzung mit dem Begriff ‚Referenzkorpus‘. Lemnitzer/Zinsmeister stellen in ihrer Korpustypologie (2015, S. 137) das Referenzkorpus dem Spezialkorpus gegenüber in Bezug auf die Beschreibungsebene „Sprachbezug“:

> Referenzkorpora sollen die Eigenschaften des dadurch repräsentierten Gegenstandes möglichst gut abdecken. Im Normalfall bedeutet Gegenstand hier eine natürliche Sprache in einer bestimmten zeitlichen Periode, zum Beispiel ‚das Deutsche des 20. Jahrhunderts‘. Referenzkorpora dienen auch als Kontrollkorpora für Untersuchungen, die sich auf Spezialkorpora beziehen und Eigenschaften der durch dieses Spezialkorpus repräsentierten Varietät untersuchen. Die Besonderheiten der untersuchten Varietät werden sichtbar, wenn man die Verteilung der zu untersuchenden Phänomene im Spezialkorpus und im Referenzkorpus vergleicht. (Lemnitzer/Zinsmeister 2015, S. 141)

Den Ausführungen ist zu entnehmen, dass Referenzkorpora eine hohe Verantwortung als Instrument der Bereitstellung von Sprachdaten für die gesamte Gruppe von mit einer natürlichen Sprache beschäftigten Wissenschaftler/-innen zukommt. Während Spezialkorpora sozusagen als Nebenprodukt des Forschungsinteresses einzelner entstehen können, hier also ein Nischendasein zwar bedauerlich, aber noch vertretbar ist, steht bei Referenzkorpora von vornherein der Community-Gedanke im Vordergrund: Das Korpus wird als Ressource für die Forschungsgemeinschaft produziert. Das bedeutet natürlich nicht, dass das Forschungsinteresse derjenigen, die mit dem Aufbau des jeweiligen Korpus betraut sind, verschwindet, der Leitgedanke einer Schaffung bestmöglicher Ansatzpunkte für externe Forschungsinteressen sollte hier aber zentral sein.

Wie aber wird ein Korpus zu einem Referenzkorpus, unter welchen Bedingungen kann ein Korpus diese Einordnung für sich beanspruchen? Die Einordnung von GiesKaNe als Beitrag zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch ist im Zusammenhang mit dem Verbund „Deutsch Diachron Digital“ zu sehen, der zur Entstehung der Referenzkorpora Altdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Mittelniederdeutsch/Niederrheinisch sowie Deutsche Inschriften führte (Dipper/Kwekkeboom 2018): „Die Referenzkorpora zu historischen Sprachstufen des Deutschen bilden die Grundlage für ein sprachstufenübergreifendes Textkorpus, das sowohl historischsynchrone als

auch diachrone Recherchemöglichkeiten bietet." (ebd., S. 95 f.). GiesKaNe soll hier quasi die Lücke zwischen dem Referenzkorpus Frühneuhochdeutsch und dem Gegenwartsdeutschen schließen. Dieser Anspruch ist aber alles andere als unproblematisch.

Mit der Anzahl der zur Verfügung stehenden potenziellen Korpustexte wächst die Komplexität der Aufgabe der Textauswahl. Während sich für das Altdeutsche die Frage der Textauswahl kaum stellt (so umfasst das Referenzkorpus Altdeutsch (ReA) die „fünf größeren Texte althochdeutscher und altsächsischer Zeit (Isidor, Tatian, Otfid, Notker und Heltland) sowie eine Vielzahl kleinerer Textdenkmäler beider Sprachstufen" (ebd., S. 96 f.), sind für die weiteren sprachhistorischen Referenzkorpora die Kriterien „Zeitraum, Sprachraum und Textart" ausschlaggebend (ebd., S. 96). Dabei handelt es sich auch um die wesentlichen Kriterien für GiesKaNe (wobei ‚Textart' hier parametrisiert wird auf der Basis von ‚Funktionalstil' und ‚Nähe-Distanz', vgl. Abschn. 3). Mit der Zunahme an potenziell nutzbaren Sprachdaten wachsen die Anforderungen an eine für den Sprachgebrauch einer Zeit repräsentative Textauswahl, der Status eines Korpus als Referenzkorpus wird dadurch schwieriger.

Neben den sprachhistorischen Referenzkorpora beansprucht das am Leibniz-Institut für Deutsche Sprache in Mannheim entwickelte und gepflegte DeReKo (= Deutsches Referenzkorpus) den Status eines Referenzkorpus. Dieser ergibt sich hier daraus, dass das IDS die größte Einrichtung zur Erforschung der deutschen Sprache ist und dass DeReKo „mit 50,6 Milliarden Wörtern (Stand: 2.2.2021) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit" bildet (DeReKo 2022). Im Gegensatz zu den sprachhistorischen Referenzkorpora müssen Korpora zur Gegenwartssprache prinzipiell Rücksicht nehmen auf die durch das Urheberrecht verursachten Einschränkungen. Das DeReKo ist deshalb als opportunistisches Korpus einzustufen, d. h., es wird fortlaufend im Wesentlichen durch solche Texte erweitert, die keine urheberrechtlichen Probleme mit sich bringen. Das führt zu einer Überrepräsentation von Pressetexten im Korpus; von einem ausgewogenen, die Gegenwartssprache repräsentativ abbildenden Referenzkorpus kann hier also keine Rede sein. GiesKaNe geht mit einer streng parametrisierten Textauswahl hingegen den Weg der sprachhistorischen Referenzkorpora, muss aber in Bezug auf die Anforderung der interoperablen Nutzung von Korpora in beide Richtungen anschlussfähig sein.

Im Gegensatz zu den älteren Sprachstufen gibt es für das Neuhochdeutsche eine weitere digitale Bereitstellung von Korpusdaten: Das Deutsche Textarchiv (DTA). Laut Geyken et al. dient das DTA „als Grundlage für ein Referenzkorpus zur Entwicklung der neuhochdeutschen Sprache" (2018, S. 219 f.). Das DTA geht folglich zurückhaltend mit dem Label Referenzkorpus um und beansprucht für

sich nur den Status einer Grundlage für ein Referenzkorpus. Vor diesem Hintergrund stellt sich für GiesKaNe die Frage, ob es tatsächlich legitim ist, nun zusätzlich zum DTA einen „Beitrag" für ein Referenzkorpus des Neuhochdeutschen anzubieten. Von einer echten Konkurrenz kann hier schon allein aus Umfangsgründen nicht gesprochen werden: Das DTA umfasst aktuell 318 Millionen Wortformen (DTA 2022), GiesKaNe strebt einen Gesamtumfang von 864.000 Wortformen an (vgl. Abschn. 3). Aufgrund der sehr unterschiedlichen Zielsetzungen der beiden Projekte kann vielmehr von einer Ergänzung gesprochen werden: Das DTA versteht sich als „Korpusaufbauprojekt" und „aktives Archiv" für die „Anlagerung weiterer Korpora" (Geyken et al 2018, S. 220) und zielt ab auf eine „möglichst vorlagengetreue Transkription historischer Quellen" sowie eine „Erfassung detailreicher Metadaten und umfangreicher Annotationen logischer und layoutbezogener Strukturen" (ebd., S. 222). Es ist also im Wesentlichen ein Instrument zur Erschließung und Bereitstellung möglichst großer Textmengen aus dem Zeitraum des Neuhochdeutschen für die linguistische Analyse. Das 129 Millionen Wortformen umfassende DTA-Kernkorpus strebt eine möglichst ausgewogene Verteilung der Domänen Zeitung, Gebrauchsliteratur, Belletristik und Wissenschaft an, für die Nutzung des DTA als „aktives Archiv" gilt aber die opportunistische Strategie der Aufnahme möglichst vieler Korpustexte. GiesKaNe, das als Beitrag zur grammatischen Fundierung eines Referenzkorpus des Neuhochdeutschen im Gegensatz zum DTA den Fokus auf die syntaktisch tiefe Annotation legt, profitiert davon, dass das DTA die Nachnutzung „in wissenschaftlichen Kontexten" ausdrücklich vorsieht (ebd., S. 221). So greift GiesKaNe im Wesentlichen auf den Bestand des DTA zurück und stellt mit einer Bereitstellung der DTA-Tokenisierung als Annotationsebene neben der GiesKaNe (GKN)-Tokenisierung (vgl. Ágel 2022) die interoperable Nutzung und damit auch die Anschlussfähigkeit an die TEI-Standards her. Einige Ergänzungen zum DTA-Bestand ergeben sich durch die Berücksichtigung von Alltagstexten im Korpusdesign von GiesKaNe. Hier nutzt GiesKaNe vor allem die Korpustexte von KAJUK (= Kasseler Junktionsprojekt, vgl. Kajuk 2009).

Für ein Korpus, das als Referenzkorpus oder zumindest als Beitrag zu einem Referenzkorpus konzipiert ist, gilt noch stärker als für sonstige Projekte zum Aufbau von Korpora: „anyone starting to undertake annotation of a corpus at a particular level should take notice of previous work which might provide a model for new work" (Leech 2004, Kap. 7). Für eine solche Orientierung an bestehenden korpuslinguistischen Ansätzen kommt korpuslinguistischen Standards eine zentrale Rolle zu.

Zu dem bereits in der Einleitung angesprochenen Problem, dass die Zementierung eines Standards eigentlich im Widerspruch zum angestrebten Fortschritt in der Wissenschaft steht, kommt allerdings als weiteres Problem hinzu: Ein

Standard kann eigentlich nur bei unveränderter Übernahme als solcher betrachtet werden. Jede Anpassung eines Standards an die spezifischen Anforderungen eines spezifischen Forschungskontexts führt dazu, dass das mit dem Standard verbundene Leitziel der maximalen Austauschbarkeit nicht mehr zu erreichen ist. Diese Problematik wird in den Abschnitten 4 und 5 mit Bezug auf TIGER und HiTs – die wichtigsten Bezugsgrößen für GiesKaNe – näher erörtert; in Abschnitt 6 erfolgt dann eine kritische Diskussion des Verhältnisses von Standard und Innovation. Den weiterführenden Überlegungen sei aber zunächst ein Überblick über den mit GiesKaNe verbundenen Ansatz vorangestellt.

3 GiesKaNe

Das Projekt „Syntaktische Grundstrukturen des Neuhochdeutschen" reagiert auf das Desiderat einer mangelnden Erforschung bzw. einer mangelnden korpusgestützten Erforschbarkeit der Syntax des Neuhochdeutschen (Ágel 2000; Elspaß 2012). Die besondere Herausforderung für einen solchen Beitrag zu einem syntaktisch erschlossenen Referenzkorpus des Neuhochdeutschen ergibt sich einerseits aus der Position des Neuhochdeutschen an der Schnittstelle von Gegenwart und Sprachgeschichte und andererseits aus der hohen Dynamik der Wandelprozesse im Untersuchungszeitraum. Für das Vorhaben ergibt sich daraus die Anforderung, eine Anschlussfähigkeit an gegenwartsbezogene und sprachgeschichtliche Forschung gleichermaßen herzustellen. Das Projekt muss folglich die objektsprachliche Ebene historisch variabler Sprachdaten ebenso berücksichtigen wie die metasprachliche Diskussion um geeignete Grammatikmodelle, wobei in Bezug auf letzteres gerade aktuelle Überlegungen zu Konvergenzen und Komplementaritäten zwischen projektionistischen und konstruktionistischen Grammatikmodellen relevant sein dürften (vgl. etwa Jacobs 2008; Welke 2011; Engelberg et al. (Hg.) 2015). Die besonderen Anforderungen an die Wandeldynamik des Neuhochdeutschen ergeben sich vor allem aus dem von Oskar Reichmann (1988) mit dem Begriff der ‚Vertikalisierung des Varietätenspektrums' beschriebenen soziokulturell bedingten Übergang von einer horizontalen zu einer vertikalen Organisation des Varietätenspektrums, d. h. von einem sozialen und räumlichen Nebeneinander von Varietäten zu einem am Leitbild einer schriftlichen Standardsprache orientierten Varietätengefüge.

Mit Blick auf die spezifischen Anforderungen in Bezug auf die Wandeldynamik des Neuhochdeutschen strebt das Projekt unter Berücksichtigung der diaphasischen, diamedialen und (teilweise) diatopischen Dimension der Variation ein ausgewogenes Korpus an, und zwar mit der folgenden Gesamtarchitektur:

Tab. 1: Geplante Gesamtstruktur von GiesKaNe

	17. Jahrhundert	18. Jahrhundert	19. Jahrhundert
Alltagstexte	Pro Jahrhundert je 72.000 Wortformen (= 6 Texte, je 2 Texte pro regionaler Raum)		
Wissenschaftstexte	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Theologie, Architektur, Geographie, Philosophie und Medizin)		
Gebrauchsliteratur	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Anstandsliteratur, Theologie, Reiseliteratur und Populärwissenschaften und 2 Texte aus dem Bereich Gesellschaft)		
Belletristik	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Reiseliteratur und Drama und je 2 Texte aus dem Bereich Prosa und Roman)		
Gesamt	288.000 Wortformen	288.000 Wortformen	288.000 Wortformen
		864.000 Wortformen	

Die Erarbeitung und Bereitstellung des Korpus erfolgt im Rahmen der Projektphasen des DFG-Langfristvorhabens. So wurde im Januar 2019 mit gieskane0.1 ein aus zwei Texten bestehendes Probekorpus über ANNIS veröffentlicht, das zunächst der Illustration des Vorhabens und Annotationsdesigns diente. Die Veröffentlichung des aus 24 Texten bestehenden und abgesehen von der Belletristik das Korpusdesign schon relativ ausgewogen abbildenden gieskane0.2 ist für den Herbst 2022 vorgesehen. Die Informationen zu den Updates können der Projekthomepage entnommen werden.

Den Anforderungen an eine Erschließung syntaktischer Grundstrukturen an der Schnittstelle von Sprachgeschichte und Gegenwart sowie projektionistischen und konstruktionistischen Grammatikmodellen begegnet das Projekt mit einem eigenen Annotationsmodell. Die zentrale theoretische Grundlage für den Ansatz bietet Vilmos Ágels Grammatische Textanalyse (2017; vgl. auch Ágel 2019). Die Innovation einer exhaustiven Annotation semantischer Rollen basiert auf dem von Ágel/Höllein (2021) veröffentlichten Ansatz. Für die syntaktisch tiefe Nominalgruppenannotation sei darüber hinaus auf Emmrich/Hennigs Ansatz zum Fokusglied (i. Dr.) verwiesen. Die korpuslinguistische Umsetzung des Annotationsmodells, insbesondere die Interaktion manueller und automatischer Arbeitsschritte, sowie die Nutzung von Verfahren des maschinellen Lernens ist in Emmrich (i. Vorb.) dokumentiert.

Das Annotationsmodell folgt dem Prinzip der Mehrebenenannotation. Herzstück ist eine eigens für das Projekt konzipierte Baumbank. Weitere Annotationsebenen umfassen ein ebenfalls neu entwickeltes POS-Tagging sowie weitere

satz- und textgrammatische Spannannotationen, die u. a. Informationen zu Parenthesen, Koordinationsellipsen und Zitation beinhalten. Als für die Frage nach dem Verhältnis von Standard und Annotation zentrale Annotationsebenen konzentriert sich der vorliegende Beitrag auf die Baumbank und das POS-Tagging.

4 TIGER vs. GiesKaNe

In den „Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora“ des DFG-Fachkollegiums 104 „Sprachwissenschaften“ (2019) wird empfohlen, Standards „mindestens als Ausgangsbasis“ heranzuziehen, sofern sie sich „für den jeweiligen Untersuchungszweck sinnvoll anwenden lassen“. Als „De-facto-Standard“ für die syntaktische Annotation wird TIGER benannt, als Standard für die morphosyntaktische Annotation STTS (2019, S. 9). Während der Status von STTS als Standard für die Wortartannotation unstrittig sein dürfte (vgl. Abschn. 5), ist die Festlegung eines Baumbank-Standards sicherlich schwieriger. Überraschenderweise befindet sich der Hinweis auf TIGER in den DFG-Empfehlungen im Abschnitt zu Tools für die Erhebung von mündlichen Korpora (was wohl daran liegt, dass in diesem Abschnitt das Themenfeld Annotation behandelt wird), obwohl TIGER anhand eines Korpus aus gegenwartssprachlichen Zeitungstexten entwickelt wurde (vgl. Eisenberg/Lezius/Smith 2005).

Die Einordnung von TIGER als de-facto-Standard lässt sich relativ einfach damit begründen, dass es in Bezug auf das Deutsche im Grunde nur zwei etablierte Baumbankmodelle gibt – neben TIGER ist das TÜBA-DZ, aus Gründen der Überschaubarkeit wird hier aber auf eine Diskussion dieses Baumbankmodells verzichtet. Aufgrund des gegenüber eindimensionaleren Annotationsebenen doch recht hohen Aufwands einer syntaktisch tiefen Annotation kommt es hier nur selten zu einem Nebeneinander von Ansätzen. Folglich kann man hier zwar von einer Tradition sprechen, aber eben nicht von einem bottom-up-Standard.

Von einer kohärenten Anwendung eines Standards über mehrere Korpora kann nur dann gesprochen werden, wenn dieser unverändert übernommen wird (vgl. dazu auch Abschn. 5). Sobald Anpassungen stattfinden, wird der Charakter als Standard geschwächt und es bedarf diverser Anstrengungen, um die jeweiligen Annotationen dennoch im Sinne eines Standards nutzen zu können (auch hierzu Abschn. 5). Dabei ist in Bezug auf die Syntax die Frage zu stellen, ob die Annahme überhaupt realistisch ist, dass ein Standard entwickelt werden kann, der für syntaktische Strukturen in sämtlichen historischen und variationellen Kontexten gleichermaßen geeignet ist. Die Gretchenfrage lautet also: Bis zu welchem Umfang an Anpassungen lohnt sich die Orientierung an einem de-facto-

Standard, ab wann ist ein Neustart zielführender? Dabei kann ein Neustart aber durchaus von den Erfahrungen bestehender Systeme profitieren, sie also im Sinne der DFG-Empfehlungen als Ausgangsbasis nutzen. GiesKaNe entspricht TIGER insofern, als Kategorien durch Knoten und Funktionen durch Kanten ausgedrückt werden und kreuzende Kanten erlaubt sind. Auch in den meisten auch teils sehr speziellen Annahmen zum Aufbau einzelner Konstituenten besteht insgesamt große Übereinstimmung. Anhand der Beispiele in Abbildung 1 und 2 seien konzeptionelle Unterschiede zwischen den beiden Baumbankansätzen aufgeführt (ohne Anspruch auf Vollständigkeit):

- TIGER basiert auf einem orthographischen Satzbegriff, d. h., die Syntaxgraphen bilden orthographische Sätze ab. GiesKaNe basiert hingegen auf einem grammatischem Satzbegriff (Ágel 2017, S. 11 f.). Folglich werden in GiesKaNe in einem Baum keine Sätze koordiniert. In TIGER hingegen bilden – wenn die Interpunktion entsprechende Satzgrenzen vorgibt – Syntaxbäume auch Satzkoordinationen ab.
- In TIGER interagiert der Baumansatz mit dem STTS-POS-Tagging. Da das STTS-Tagset ein morphosyntaktisches Wortarttagging bereitstellt, das syntaktische Informationen wie etwa die Position in der Linearstruktur und teilweise auch Angaben zur Funktion wie bspw. zum attributiven Gebrauch von Adjektiven enthält, verzichtet der Baumbankansatz auf eine Ausdifferenzierung der terminalen Kanten in Wortgruppen:

  > Eine NP besteht zunächst aus einer Reihe von pronominalen, substantivischen und adjektivischen Kernelementen (NP kernel elements, NK). Ihre genauere Unterteilung kann aufgrund der Part-of-Speech bzw. kategorialen Information vorgenommen werden, so daß sich eine Unterscheidung auf der Ebene der Funktionslabels erübrigt. (Albert et al. 2003, S. 9)

  GiesKaNe dagegen setzt auf ein modulares System der Mehrebenenannotation, in dem die Baumbank die alleinig verantwortliche Annotationsebene für die Syntax ist.
- Auf Satzebene besteht der zentrale grammatiktheoretische Unterschied darin, dass in TIGER das finite Verb zentral für den Satz ist (Sätze werden hier auch als „Phrasen mit finitem Verb“ definiert, vgl. Albert et al. 2003, S. 48), in der Konsequenz wird es als Kopf des Satzes annotiert. In GiesKaNe hingegen ist das Prädikat das Zentrum des Satzes. Die Konsequenz bei TIGER ist, dass nicht-finite Teile von Sätzen als Verbalphrasen annotiert werden, die neben dem nicht-finiten Verb auch die Satzglieder außer dem Subjekt enthalten. Diese Festlegung hängt offenbar damit zusammen, dass TIGER auf NEGRA basiert, ein in Saarbrücken erstelltes Korpus deutscher Zeitungstexte (vgl. Eisen-

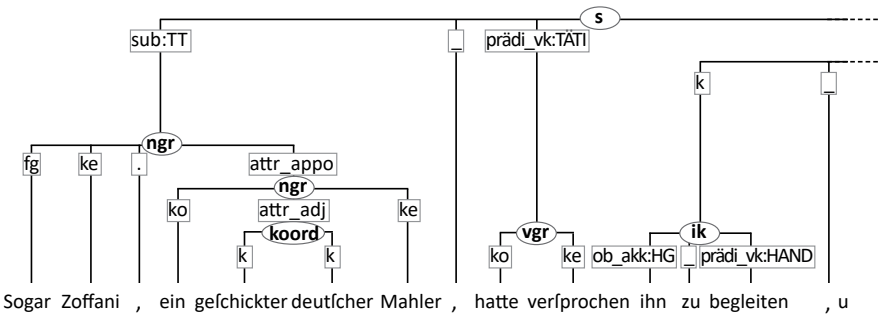

Abb. 1: Baumbeispiel GiesKaNe (Bauernleben, 17. Jh.)

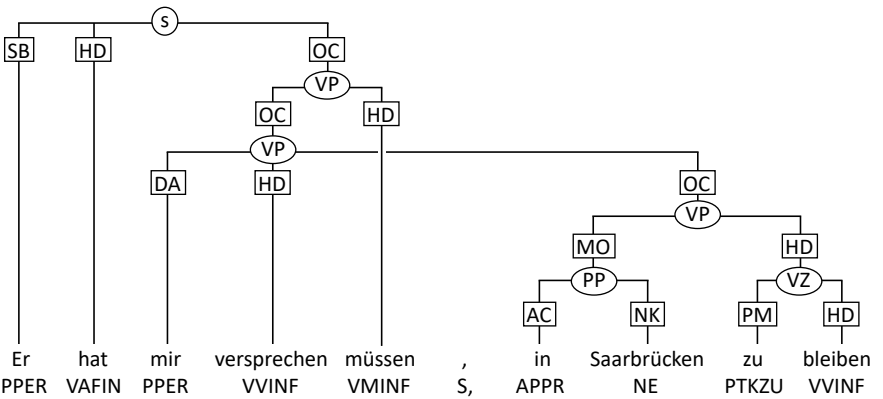

Abb. 2: Baumbeispiele TIGER (Albert et al. 2003, S. 50, 69)

berg/Lezius/Smith 2005, S. 81). Sie weicht jedenfalls ab von der von Eisenberg in seiner Grammatik vertretenen Konstituentenstrukturgrammatik. Eisenberg spricht sich dort am Beispiel der Diskussion mehrerer Analysemodelle des Satzes *Karl will Bier holen* gegen die in TIGER praktizierte Variante aus mit dem Argument, dass diese „zwar die Objekt-Funktion von Bier angemessen erfass[en würde], nicht aber die syntaktischen Beziehungen zwischen *will* und *holen* sowie die zwischen *Karl* und *holen*" (2020, S. 98). In der Eisenberg'schen Konstituentenstrukturgrammatik werden folglich aus verschiedenen Verben bestehende Verbalkomplexe einheitlich als Verbgruppen erfasst (unabhängig davon, welche Art von Spezialverb das Vollverb begleitet) und Infinitivkonstruktionen als Infinitivgruppen in der Konstituentenstruktur analog zu Nebensätzen verortet. GiesKaNe folgt in diesem Sinne der Eisenberg'schen Konstituentenstrukturgrammatik.

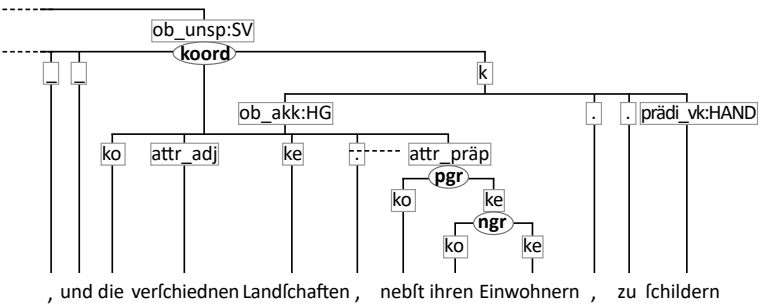

Abb. 1 (Fortsetzung)

- Ein weiterer zentraler Unterschied der Annotation satzgrammatischer Grundstrukturen besteht in der Annotation von semantischen Klassen von Prädikaten und Satzgliedern in GiesKaNe. Als Grundtypen semantischer Klassen werden hier Prädikatsklassen, Adverbialklassen und semantische Rollen annotiert (vgl. Annotationshandbuch Ágel/Hennig 2022 sowie Ágel/Höllein 2021). TIGER enthält keine Angaben zur Satzsemantik.
- GiesKaNe unterscheidet insgesamt stärker zwischen satz- und wortgruppengrammatischen Strukturen als TIGER. So kommen die funktionalen Werte Modifizierer und Objekt in TIGER sowohl als Werte für Satz- als auch für Wortgruppenfunktionen in Frage (Albert et al. 2003, S. 27 f.). Dabei kommt es in TIGER bei deverbalen nominalen Kernen zur Annotation von Objekten (Objektsätze und Präpositionalobjekte). GiesKaNe entgegen operiert in diesen Fällen mit einer Attributannotation. Die Vergabe des Werts Modifizierer in TIGER ist dagegen nicht an spezifische Bedingungen dieser Art gebunden, sie wird vielmehr gleichermaßen für verschiedene satz- und wortgruppengrammatische Typen der Modifikation genutzt (Adverbiale, Fokus- und Intensitätspartikeln in Nominal- und Adjektivgruppen, Erweiterungen von Adjektiv- und Partizipialattributen).
- Auch in Bezug auf die interne Struktur von Wortgruppen kann GiesKaNe insgesamt eine stärkere Nähe zum Eisenberg'schen Konstituentenstrukturformat attestiert werden als TIGER. Wie bereits erwähnt, verzichtet TIGER auf eine Ausdifferenzierung von Wortgruppengliedern (im Sinne von Ágel 2017, S. 23). In GiesKaNe kommen als Wortgruppenglieder Köpfe, Kerne und Attribute in Frage (in gewisser Hinsicht auch Fokusglieder, vgl. Emmrich/Hennig i. Dr.). Als interne Struktur von Präpositionalgruppen nimmt GiesKaNe in Anlehnung an Eisenbergs Konstituentenstrukturgrammatik eine rekursive Gruppenstruktur an, d. h., als Kern der Präpositionalgruppe kommt hier eine Nominalgruppe in Frage. TIGER hingegen sieht eine flache Struktur von Präpositionalgruppen vor.

Der Stellenwert der einzelnen Unterschiede für die Entscheidung für GiesKaNe im vorliegenden Projekt muss differenziert betrachtet werden: Das Ausgehen von Interpunktion als Kriterium für die syntaktische Einheitenbildung ist für eine Erstellung eines syntaktisch annotierten Korpus des Neuhochdeutschen tatsächlich auszuschließen. Einerseits ist die Grammatikalisierung der Interpunktion als zuverlässiger Indikator für grammatische Verhältnisse ja gerade erst Gegenstand der im Untersuchungszeitraum liegenden Standardisierungsprozesse, was andererseits gerade für die in der Korpusarchitektur berücksichtigten Alltagstexte in teilweise noch erheblich höherem Maße gilt. Was diejenigen Bereiche anbelangt, die in GiesKaNe detaillierter abgebildet sind, so ist hier hingegen zu konstatieren, dass man darin einerseits einen Mehrwert sehen kann, dass aber andererseits eine Erschließung syntaktischer Grundstrukturen des Neuhochdeutschen prinzipiell auch in weniger detaillierter Form möglich wäre. Vor allem aber möchten wir hier keine Diskussion darüber führen, welcher grammatiktheoretische Ansatz besser für die Arbeit mit den sprachhistorischen Daten geeignet ist. So sei an dieser Stelle ausdrücklich betont, dass mit der Dokumentation wesentlicher Unterschiede der Baumbankansätze von TIGER und GiesKaNe keine Wertung verbunden sein soll. Es wird vielmehr deutlich, dass in die Konzeption einer Baumbank unweigerlich eine Vielzahl an grammatiktheoretischen Grundsatzentscheidungen eingehen. Bekanntlich ist die Präferenz für eine grammatiktheoretische Erklärung eine Frage der Schulenbildung und diese zu bewerten, ist nicht Anliegen des vorliegenden Beitrags. Für die Frage der Festlegung auf ein Modell für die korpuslinguistische Erschließung von syntaktischen Strukturen im Format einer Baumbank dürfte neben der natürlich grundlegenden Frage nach der Eignung des Modells für das mit der Korpuserschließung verknüpfte Forschungsinteresse die eher pragmatische Frage der Umsetzbarkeit mit den zur Verfügung stehenden Ressourcen zentral sein. Bei einem so komplexen System, wie es einer Baumbank zugrunde liegt, ist eine punktuelle Anpassung an veränderte Kontexte sicherlich schwierig, zumal bei punktuellen Eingriffen immer auch mit Konsequenzen für andere, eigentlich nicht dem Anpassungsinteresse unterliegenden, Bestandteile des Gesamtsystems gerechnet werden muss. Kurzum: Im Sinne einer Kosten-Nutzen-Rechnung sowie natürlich auch auf der Basis eigener grammatiktheoretischer Überzeugungen wurde hier der Neustart als der geeignetere Weg angesehen (vgl. auch Abschn. 6).

Ob sich TIGER tatsächlich längerfristig als Standard für die syntaktische Annotation deutschsprachiger Korpora durchsetzen wird, kann hier nicht antizipiert werden. Da eine Vergleichbarkeit bzw. interoperable Nutzung verschiedener syntaktisch annotierter Korpora selbstverständlich anzustreben ist, planen wir für GiesKaNe die Ergänzung einer Annotationsebene mit einem auf der Basis

maschinellen Lernens erstellten TIGER-Parsers. Der Gedanke, dass dem Standard-vs.-Innovation-Dilemma mit einem solchen Ansatz sinnvoll begegnet werden kann, sei im Folgenden anhand der POS-Standards STTS und HiTs erläutert.

5 Zwischen Standard und Innovation: HiTs

Das Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) ist sicherlich der de-facto-Standard der germanistischen Korpuslinguistik schlechthin. STTS, dessen Ziel ganz im Sinne der Standardorientierung in einer „weitgehende[n] Übereinstimmung der Korpus Annotation [...], die die gegenseitige Nutzung bereits durchgeführter Korpusarbeit ohne umständliche Anpassung unterschiedlicher Tagsets" beinhaltet, besteht (ebd., S. 3), wird im Grunde genommen flächendeckend für die Erschließung von Wortartkategorien genutzt. Die Anhebung zu einem quasi-de-jure-Standard durch die DFG-Empfehlungen hat in diesem Fall also eine deutlich solidere Grundlage als im Falle von TIGER. Anpassungen erwiesen sich jedoch in bestimmten Kontexten dennoch als notwendig – zu nennen wäre hier „STTS 2.0" für die gesprochene Sprache (Westphal et al. 2017) sowie „HiTS" für historische Sprachkorpora. Es ist sicherlich kein Zufall, dass gerade diese beiden Anwendungsfelder eine Anpassung des Standards als notwendig erscheinen ließen – offenbar zielt STTS zunächst auf die geschriebene Standardsprache der Gegenwart ab.

Für die Diskussion des Umgangs mit Standards in unserem Kontext ist HiTS einschlägig. Indem HiTS innerhalb der community der Sprachhistoriker/-innen für die Bedarfe der Annotation historischer Korpora entwickelt wurde, kann es wiederum als ein Beispiel für einen bottom-up-Standard angesehen werden.

> HiTS orientiert sich am „Stuttgart-Tübingen Tagset" (STTS, Schiller et al., 1999), dem Standardtagset für nhd. Korpora, und übernimmt — neben einer ganzen Reihe von Tags — auch das hierarchische Design der Tagnamen. Ursprünglich sollte das Tagset komplett auf STTS aufbauen und dieses lediglich um einige neue Tags erweitern. Es stellte sich jedoch heraus, dass neben einigen notwendigen feineren Unterscheidungen (z. B. bei den Pronominaladverbien) auch die Tagnamen des STTS nicht immer geeignet schienen. (Dipper et al. 2013, S. 85)

Eine wesentliche Anpassung besteht auch in der Festlegung der Anwendung des Tagsets auf die Beleg- und Lemmaebene: „In HiTS wird die Wortart einer jeden Wortform zweifach annotiert, und zwar zum einen mit Blick auf das Lemma und zum anderen mit Blick auf den konkreten Beleg, also der Verwendung einer Wort-

form in einem spezifischen Kontext.“ (ebd., S. 92). Dadurch wird deutlich, dass es bei der Frage nach der Etablierung eines Standards in der Korpuslinguistik keineswegs nur um Standardtags geht, sondern auch um die Frage, was eigentlich mit den Tags annotiert wird. So kann von einer vollständigen Standardorientierung nur dann gesprochen werden, wenn

- das Tagset uneingeschränkt, also ohne Anpassungen übernommen wird;
- die Kriterien der Annotation identisch sind;
- die Tags auf die gleiche Annotationsebene (bspw. Tokenebene, Lemmaebene) bezogen werden.

Auch innerhalb der Anwendung von HiTS in den verschiedenen sprachhistorischen Referenzkorpora kommt es zu Unterschieden, wie die folgende Übersicht anhand von Tags zum Adjektv anschaulich illustriert:

STTS	**DDDTS**	**DDDTS-HIPKON**	**HiTS**	**HiNTS**	**Beschreibung**
ADJA	ADJ	ADJ	ADJA	ADJA	atributives Adjetkiv (oder eliptisch)
ADJD	ADJD	ADJD	ADJD	ADJD	(adverbiales oder) prädikatives Adjektiv
	ADJE	ADJE			Adjektiv, attributiv, Teil eines Eigennamens
	ADJN	ADJN	ADJN	ADJN	Adjektiv, attributiv, nachgestellt
	ADJNE	ADJNE			Adjektiv, attributiv, nachgestellt, Teil eines Eigennamens
	ADJO	ADJO			Adjektiv, ordinal, attributiv
	ADJON	ADJON			Adjektiv, ordinal, attributiv, nachgestellt
	ADJOS	ADJOS			Adjektiv, ordinal, substantiviert
	ADJS	ADJS			Adjektiv, substantiviert
			ADJS	ADJS	Adjektiv, substituierend
				ADJV	Adjektiv, adverbial
				ADJ...	Adjektivische Ordinalzahl
	ADJOE				Adjektiv, ordinal, attributiv, vorangestellt oder elliptisch, Teil eines Eigennamens

Abb. 3: STTS, HiTS und weitere Anpassungen im Vergleich (Odebrecht 2017, S. 14): DDDTS = Deutsch Diachron Tagset (Altdeutsch); HIPKON = Historisches Predigtenkorpus; HiNTS = Tagset für Mittelniederdeutsch (Barteld et al. 2018)

Vor diesem Hintergrund ist durchaus die Frage zu stellen, ob in Bezug auf HiTS tatsächlich von einem Standard gesprochen werden kann. Eine gemeinsame Basis ist vorhanden, für die interoperable Nutzbarkeit der Korpora sind aber weitere Anstrengungen vonnöten.

Dass GiesKaNe zunächst nicht auf HiTS zurückgreift und mit einem eigenen Tagset für die Wortartannotation arbeitet, kann damit begründet werden, dass

GiesKaNe insgesamt auf eine modularere Annotation im Modell der Mehrebenenannotation setzt. Während STTS, HiTS und die verwandten Tagsets Tags wie ADJA, ADJD, ADJN und ADJS enthalten, die als fusionierende Tags Informationen zur Wortart sowie zur syntaktischen Funktion des Worts im Kontext sowie zu Stellungseigenschaften enthalten, beschränkt GiesKaNe die Wortartannotation auf die Annotation von Wortarten im engeren Sinne und nimmt keine syntaktischen Eigenschaften in die POS-Annotation auf, da die genannten syntaktischen Eigenschaften in der Baumbank erfasst sind: Das Wortarttagging ist damit sozusagen von dieser Aufgabe entbunden. Mit dieser stärker modularen Organisation ist eine größere Flexibilität gegeben, GiesKaNe setzt also auf die vielfältigen Kombinationsmöglichkeiten der Annotationsebenen.

6 Standard oder Innovation

Bevor abschließend mit einer Studie zur Anwendung des maschinellen Lernens bei bereits bestehenden manuellen Annotationen eine Lösung für den angesprochenen Konflikt zwischen Standard und Innovation am Beispiel von HiTS vorgestellt wird, soll hier noch einmal nachvollzogen werden, wieso es überhaupt zu diesem Konflikt kommt und welche weiteren Dimensionen der Arbeit mit Annotationen hierbei berücksichtigt werden müssen. Denn grundsätzlich lassen sich Annotationen hinsichtlich ganz verschiedener Faktoren verorten: Manuelle Annotationen entstehen prinzipiell in Forschungsprojekten, die ein Forschungsinteresse verfolgen. Eine Analyse von Hand kann Unbekanntes oder Abweichendes beschreiben und Probleme offenlegen, ist allerdings auch zeit- und kostenintensiv. Entsprechend muss ggf. das projektinterne Forschungsinteresse als Ziel der Arbeit in Bezug auf eine Verpflichtung gegenüber der Forschungsgemeinschaft relativiert werden: Die aufwendige Arbeit ist vor allem dann gerechtfertigt, wenn das Produkt auch eine Ressource für die Forschungsgemeinschaft darstellt. Dabei ist schon die Frage, wie gut selbstgewählte Mittel ein Forschungsvorhaben ermöglichen, nicht vorab leicht zu beantworten, und eine Antwort wird umso schwerer, wenn mögliche Interessen der Gemeinschaft antizipiert werden müssen. Bezogen auf das Korpus den Umfang der Annotationen zu steigern und so verschiedenen Interessen gerecht zu werden, steht dann im Konflikt zum Aufwand und den durch Zeit und Kosten gesetzten Grenzen oder aber zu der als Ausgangspunkt der Überlegungen gewählten Qualität manueller Annotationen. Problemorientiertes Arbeiten wird erschwert, wenn sich der Umfang der Analysen erhöht. Natürlich kann die mehrfache Annotation einer Textstelle auch als Chance begriffen werden. Das ändert aber nichts am Ausgangsproblem der durch Zeit und Kosten

gesetzten Grenzen. Maschinelle Verfahren wiederum können einerseits nicht immer als Alternative zu manueller Annotation betrachtet werden und sind andererseits auf manuelle Annotationen angewiesen – jedenfalls im Bereich des maschinellen Lernens.

Innerhalb dieser Dimensionen ist Standardisierung im Bereich von Annotationen zu diskutieren: Als Ressource der Forschungsgemeinschaft muss das Korpus möglichst leicht zugänglich sein und in die bestehende Infrastruktur eingebunden werden. Beides kann durch Standardisierung erreicht werden. Demgegenüber erscheint der Gedanke, ein Forschungssinteresse mit seinem Anspruch an Innovation durch standardisierte Mittel zu verfolgen, problematisch. Gerade das Potenzial manueller Annotationen im Sinne eines problemorientierten Arbeitens kann nur eingeschränkt oder gar nicht genutzt werden, wenn die Analyseentscheidungen bereits definiert sind. Problematisch ist weniger die Verwendung der Knoten- und Kantenlabel oder des Tagsets an sich, sondern die dahinterstehenden Abgrenzungskriterien, Tests, Kategorienbildungen, die einheitlich angewendet werden müssen, um übereinstimmende Annotationen vorzunehmen. Eine vermittelnde Perspektive, bei der ein bestehendes Annotationsschema grundsätzlich übernommen, aber punktuell abgewandelt wird, könnte möglicherweise beiden Perspektiven auf das Korpus nicht gerecht werden.

Schon das Verhältnis von eingesetztem Mittel zu Forschungsinteresse bzw. zwischen Annotationen und Forschungsinteresse ist mitunter problematisch, wenn – wie in unserem Fall – ein Forschungsinteresse im Bereich des Neuhochdeutschen besteht und als Mittel Annotationen für entsprechende Texte vorgenommen werden. Grundsätzlich ist jede Forschung wohl weder in Hinblick auf Theorien unvoreingenommen, noch wird sie trotz anderer Datenlage an vorherigen Annahmen festhalten (vgl. Wegera 2013). Erstere würden u.E. den bisherigen Diskurs ignorieren und letzteres die Spielregeln. Die historische Sprachwissenschaft war schon immer auf Daten angewiesen und der Aufbau eines Korpus zur Syntax des Neuhochdeutschen ist ohne Annotationen kaum vorstellbar. Daher ergibt sich bezogen auf unser Vorhaben das Problem, dass Syntax erforscht werden soll, dazu Annotationen vorgenommen werden und diese auf syntaktischen Analysen beruhen – obwohl ja streng genommen erst das fertige Korpus die Datengrundlage für syntaktische Analysen bieten soll. Bei der Vornahme manueller Annotationen müssen Probleme erkannt und vergleichend auf der Basis der nicht annotierten Texte und bestehender Korpora betrachtet werden. Somit ist gerade der Prozess der Korpuserstellung für das Forschungsprojekt zentral, wenn hier die Schritte zur Erforschung des Gegenstands vorgenommen werden. Das unterstreicht die Bedeutung des problemorientierten manuellen Annotierens und den Konflikt, der zur Anwendung eines Standards bestehen kann – aber auch zum Verhältnis von Arbeitszeit und Umfang der Annotationen. Die

Rolle des fertigen oder jeweils fertigen Korpus wird dadurch nicht gemindert. Sie besteht vielmehr darin, einen Überblick zu erhalten und eine Datenbasis für die Bearbeitung aufbauender Fragestellungen bereitzustellen. In jedem Fall darf die anschauliche digitale Erscheinungsform des fertigen Korpus nicht vergessen lassen, dass dieses im Grunde auf der eigenen Anreicherung mit Informationen basiert: Man findet sonst – so bringt es Wegera (2013) auf den Punkt – die Ostereier dort, wo man sie selbst versteckt hat, und ist darüber womöglich noch überrascht.

Im Bemühen um Theorieneutralität oder allein wegen einer sicherlich bestehenden, aber undefinierbaren Forschungslücke wäre ein Verzicht auf Annotationen, wie angesprochen, ein radikaler Schritt, weil Annotationen, wie Gries/Berez (2017) festhalten, nur einen Mehrwert darstellen: Man muss sich nicht auf Annotationen verlassen und kann sie letztlich auch gänzlich ignorieren. Überspitzt gesagt fände diese Perspektive ihre Grenzen in einer einfachen Kosten-Nutzenrechnung, wenn die verlässlichste und meistgenutzte Ebene einer Baumbank die Tokenebene wäre, weil Annotationen unter speziellen, wenig anschlussfähigen theoretischen Annahmen gemacht werden oder aber nicht die notwendige Qualität aufweisen und nicht verlässlich sind. Wenn Annotationen vorgenommen werden, müssen die bisher diskutierten Faktoren berücksichtigt werden, weil dem hohen Aufwand auch ein hoher Nutzen gegenüberstehen muss. Annotationen als Mehrwert zu betrachten und mehrere unabhängige Annotationsebenen anzubieten – also etwa einen Standard als Alternative zum gewählten Annotationsschema –, scheint trotz bestehender Einwände unter den gegebenen technischen Voraussetzungen der Mehrebenenannotation ein zielführender Ansatz, wenn der Aufwand minimiert und die notwendige Qualität gewährleistet werden kann. Ein Ansatz könnte maschinelles Lernen auf der Basis bestehender Annotationen sein, um einen Standard als alternative Annotationsebene anzubieten. Bevor dieser Ansatz im folgenden Abschnitt mit einer praktischen Studie vorgestellt wird, soll abschließend noch diskutiert werden, ob bestehende Annotationsschemata eine Alternative zur Entwicklung eines Annotationsschemas im Rahmen eines Forschungsvorhabens darstellen können.

Mit der Verwendung von bestehenden Annotationsschemata wie STTS, HiTS oder TIGER würde der Perspektive Korpus als Gemeinschaftsressource Rechnung getragen werden, wobei dann allerdings das projektinterne Forschungsinteresse zurückgestellt werden müsste, gleichzeitig aber auch Usability und Vergleichbarkeit verbessert werden könnten. Auch innerhalb des auf ein bestimmtes Forschungsinteresse ausgerichteten Projekts könnte man so Korpuserstellung und -nutzung bis zu einem gewissen Grad entkoppeln und so das Ostereiproblem begrenzen. Zudem würde es der wissenschaftlichen Praxis entsprechen, wenn ein Erkenntnisinteresse auf das bestehende Wissen aufbaut. Hier aber muss der

Begriff des Standards erneut betrachtet und es muss die Frage gestellt werden, was überhaupt weshalb als Standard begriffen werden kann. Auch wenn in Bezug auf Annotationsschemata der Begriff Standard im Sinne eines community-driven Standards (Leech 2004) verstanden werden kann, stellt sich die Frage, ob darin ein durch Kritik und Übernahme gefestigtes Wissen, wie es der wissenschaftliche Diskurs hervorbringt, zum Ausdruck kommt und inwiefern eine etablierte Praxis gegeben ist. Der Aufbau eines Annotationsschemas stellt – gerade bei Baumbanken – eine komplexe Aufgabe dar, an der mehrere Personen oder Gruppen beteiligt sind; daher gehen sie in der Regel aus Projekten hervor. Diese sind aber eben kosten- und zeitintensiv und entsprechend selten, sodass die Etablierung als community-driven Standard schon aus der Seltenheit selbst folgt – so wenigstens die Argumentation von Pustejovsky/Stubbs (2012) zu Standards im Bereich der Auszeichnungsformate. Daher kommt es seltener zu einem Nebeneinander von Ansätzen; unterschiedliche Erfahrungstraditionen können sich nur langsam oder gar nicht entwickeln bzw. weiterentwickeln; Kritik erfolgt gar nicht oder verzögert. Das zeigt sich auch daran, dass kritische Auseinandersetzungen mit entsprechenden Standards selten sind: etwa die Entwicklung von HiTS (Dipper et al. 2013) für historische und STTS-2.0 (Westpfahl et al. 2017) für gesprochensprachliche Texte auf der Basis des STTS (Schiller et al. 1999;vgl. Abschn. 5). Und diese Schritte setzen den Etablierungsprozess ja erst in Gang.

Ein weiterer Aspekt betrifft das Verhältnis von Abwandlung zu Usability und Vergleichbarkeit. Zwar muss für einen community-driven Standard kein Entweder-Oder gelten. Um Usability und Vergleichbarkeit aber möglichst hoch zu erhalten, müssten Umfang und Anzahl der Änderungen möglichst gering gehalten werden. Bei komplexen Systemen wie den Annotationsschemata für Baumbanken stellt sich jedoch die Frage, ob punktuelle Eingriffe und Änderungen möglich sind. Wie bei einer Grammatik kann man nicht einfach Konzepte ändern, hinzufügen oder tilgen, ohne dass andere Bereiche davon betroffen wären, und so führt jede Änderung zu weiteren Änderungen und so verändert sich das System, was dann wiederum zu immer größerer Beeinträchtigung von Usability und Vergleichbarkeit führt. Da Änderungen bei Standards nicht unproblematisch sind, stellt sich die Frage, wie sie sich zum Streben der Wissenschaft nach neuer Erkenntnis verhalten. Während Standards auf Übernahme angewiesen sind, ist das Streben nach neuer Erkenntnis charakteristisch für die Wissenschaft. Würde man etwa Standards strikt anwenden, wäre jedes neue Korpus – wie bereits angesprochen – als quantitative Erweiterung zu betrachten. Bezogen auf das GiesKaNe-Korpus, könnte mit der Verwendung von TIGER nicht mehr die Segmentierung, Kategorisierung und Hierarchisierung in Texten geändert werden, sondern nur die Auswahl derselben unter Aspekten wie zeitlicher oder bspw. nähe-und-distanzsprachlicher Variation. Dabei wäre das angesprochene Ostereiproblem unter

Umständen nicht gelöst, sondern verstärkt, weil man bei der Anwendung von Standards dann in vielen Korpora das findet und bestätigt sieht, was durch die wiederholte Anwendung eines Annotationsschemas annotiert wurde. Letztlich stellt sich auch die Frage der Eignung. Denn einer Vielzahl von Forschungsinteressen kann unmöglich eine geeignete Menge an Standards gegenüberstehen. Wenn TIGER etwa als Standard für Baumbanken empfohlen wird, zeigt schon die Abänderung des STTS im Sinne von HiTS und STTS-2.0, dass TIGER nicht als Standard für historische Baumbanken gelten kann: Wenn bereits flache POS-Tagging-Ansätze anpassungsbedürftig sind, ist kaum zu erwarten, dass das anhand von gegenwartssprachlichen Zeitungstexten entwickelte TIGER-Schema, das als Baumbankansatz eine größere Tiefe und Komplexität aufweist als ein POS-Tagging, den Anforderungen historischer Texte uneingeschränkt gerecht werden kann. Hinzu kommt, dass TIGER selbst auf die Verzahnung mit STTS setzt, das nicht als POS-Standard für sprachhistorische Korpora gelten kann (vgl. Abschn. 5). Grundsätzlich ist also theoretisch wie praktisch der Begriff des Standards und die Anwendung von Standards im Rahmen der Forschung problematisch, was ihre Vorteile in Bezug auf Vergleichbarkeit und Usability allerdings nicht in Frage stellt.

Zurückgestellt wurde in der bisherigen Diskussion der Gedanke, dass mehr statt weniger Annotationen ein möglicher Lösungsansatz sein könnte. Im Sinne der Perspektive von Annotationen als einfachem Mehrwert könnte man Annotationen nach einem Standard einfach neben anderen, innovativen Annotationen realisieren: Annotationen könnten flexibel an das Forschungsinteresse angepasst werden, der für das Forschungsinteresse zentrale Aspekt des problemorientierten Annotierens würde kaum beeinträchtigt, durch ein Nebeneinander von Innovation und Standard stünde beiden Seiten ein Korrektiv zur Verfügung, Standards könnten durch Kritik und Übernahme weiter gefestigt werden, Usability und Vergleichbarkeit wären unter Berücksichtigung der anderen Faktoren bestmöglich gewährleistet und würden zudem den Zugang zum unbekannten Annotationsschema erleichtern. Kritische Größe ist hier der Aufwand bzw. die Qualität der Annotationen. Ein Lösungsansatz könnte u.E. der Einsatz von maschinellem Lernen sein. Maschinelle Verfahren gehören zum Werkzeugkasten bei der Erstellung von Korpora und auch maschinelles Lernen bzw. Deep Learning sind bewährte Mittel des Korpusaufbaus. Unser Vorschlag fokussiert dabei die Wiederverwertung der manuellen Annotationen mit ihrer hohen Qualität und den genauen Informationen zum Kontext. Es geht folglich darum zu zeigen, wie gut ein in manueller Annotation angewendetes Tagset auf diese Weise zur Ableitung eines anderen Tagsets genutzt werden kann – in unserem Fall, wie gut ein Standard wie HiTS auf der Basis der Annotationen in GiesKaNe ergänzt werden kann.

7 Wiederverwertung manueller Annotationen durch maschinelles Lernen

Der Grundgedanke ist dabei, dass ein Tagger oder Parser also nicht wie üblich bei Null anfangen muss, sondern bestehende Annotationen, die letztlich auch nur gleiche oder vergleichbare Merkmale der Sprache erfassen, nutzt. Üblicherweise greift ein Tagger etwa auf die Wortform im Kontext einer Eingabesequenz wie einem Satz zurück. Je nach Sprachstufe, Konzeption und Textsorte können womöglich *word embeddings* unterschiedlicher Art eingebunden werden. Sind diese Tokensequenzen bereits annotiert, stehen dem Tagger noch abstraktere Kategorien als Merkmale zur Verfügung, die sozusagen als hochwertige Eingabe-Merkmale eine noch genauere Differenzierung ermöglichen. Der Tagger muss quasi nur das Übersetzen lernen. Das soll abschließend durch die Anwendung von HiTS in GiesKaNe veranschaulicht werden. Der Tagger basiert auf einem CRF-Modell (Lafferty/McCallum/Pereira 2001). Annotationen der Textabschnitte in GiesKaNe wurden um HiTS-Tags erweitert. GiesKaNe-Annotationen wie Wortart, syntaktische Funktion, Wortart und syntaktische Funktion des vorherigen und nächsten Wortes dienen dann als Eingabe-Werte/Features.

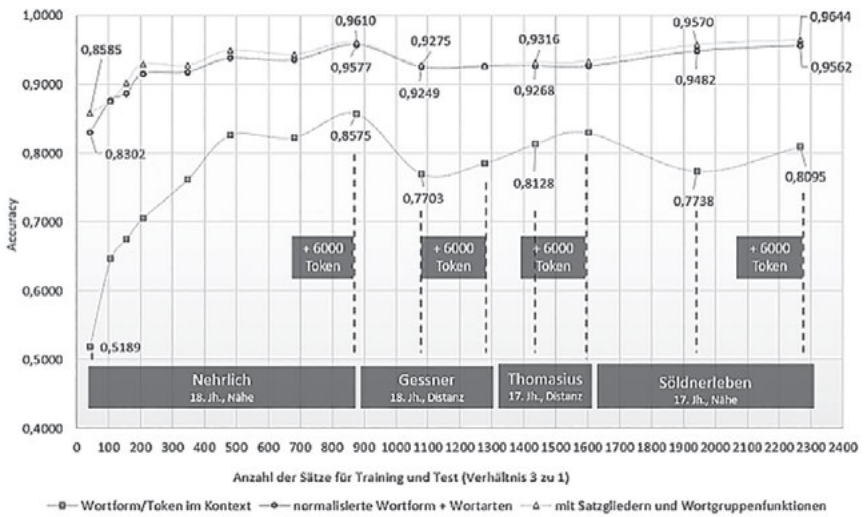

Abb. 4: HiTS-Tagger in GiesKaNe, 3 Modelle

Auf der y-Achse wird die Genauigkeit des Tagger-Modells abgebildet. Auf der x-Achse die Anzahl der für Training und Test des Modells verwendeten Sätze/

Eingabesequenzen. Die viereckigen Datenpunkte zeigen ein einfaches Tagger-Modell auf der Basis der Textoberfläche im Kontext der Eingabesequenz, die runden Datenpunkte ein Modell auf der Basis normalisierter Wortformen und der Wortartentags nach dem GiesKaNe-Tagset zum Vergleich, weil eine Baumbank nicht immer vorausgesetzt werden kann. Im Modell mit den dreieckigen Datenpunkten wurden dann noch zusätzlich syntaktische Informationen aus der Baumbank genutzt. Während die auf Annotationen aufbauenden Modelle schon bei 200 Sätzen (ca. 2.000 Token) eine Genauigkeit von über 90% erreichen, liegt der ‚einfache', auf der Textoberfläche aufbauende Tagger hier bei gerade einmal 70%. Bei rund 2.300 Sätzen und etwa 24.000 Token erreichen die annotationsbasierten Modelle eine Genauigkeit von 95 bzw. 96%. Ein einfacher Tagger liegt hier – zum Vergleich – bei gerade einmal 81%. Auffällig sind zudem die leichten Schwankungen der ersten beiden Modelle und die starken Schwankungen des einfachen Taggers, die hier mit den Grenzen der zum Training genutzten Texte übereinstimmen und daher durch textspezifische Besonderheiten erklärt werden können. Gerade das einfache Tagger-Modell könnte gegenüber diesen textspezifischen Besonderheiten – gerade im Bereich konzeptioneller Mündlichkeit – anfällig für Probleme in Zusammenhang mit der Variation von Wortformen und Konstruktionen im Kontext sein, während sich diese Faktoren auf die annotationsbasierten Modelle nach dem Aufbau eines Grundumfangs an Daten möglicherweise weniger auswirken. Die abstrakteren Wortartanalysen und die normalisierten Wortformen würden dann als Abstraktionen diese Faktoren womöglich schnell und beständig ausgleichen. In der durch Abbildung 4 veranschaulichten Studie wurden die Trainingsdaten nicht wie üblich gemischt, um den Effekt textspezifischer Besonderheiten auch in dieser kleinen Studie veranschaulichen zu können. Mischt man die Eingabesequenzen, erreicht das einfache Modell eine Genauigkeit von 88% und die Genauigkeit ließe sich mit den angesprochenen Verfeinerungsschritten weiter steigern.

Entscheidend ist letztlich aber, dass der Aufwand bei den auf Annotationen aufbauenden Modellen in Bezug auf die parallel zu annotierenden Texte nicht nur relativ zu der in unserem Projekt angestrebten Menge an Texten überschaubar ist. Der Ansatz ließe sich also auch auf kleinere Projekte übertragen, wenn sich die Genauigkeit schon früh bei über 95% stabilisiert. Auch die letztlich erreichte Genauigkeit von über 96% nach dem Training auf der Basis von ca. 18.000 bzw. 24.000 Token (2,8% des Gesamtumfangs des Projekts) liegt nur wenige Prozentpunkte unter dem Bereich oder sogar in dem Bereich, der für die manuelle Annotation in IAA-Studien zu vergleichbaren Tagsets angegeben wird: STTS/NEGRA (gegenwartssprachliche Zeitungstexte): 98,57% (Brants 2000), HiNTS/ReN (Mittelniederdeutsch/Niederrheinisch): 94,33% (Barteld et al. 2018), STTS-EMG/GerManC (Neuhochdeutsch): 91,6% (Scheible et al. 2011). Für die

Wortartebene in GiesKaNe und unser Tagset (Höllein/Lotzow 2019) liegt der Wert bei 95,8%. Daher scheint auch eine unkorrigierte Anwendung auf das Korpus möglich. Die Perspektive von Annotationen als Mehrwert kann vor diesem Hintergrund als Lösung für viele besprochene Probleme angenommen werden, weil die Bedenken bezüglich des Aufwands und der Qualität auf diese Weise deutlich gemindert werden.

Sicherlich bleibt im Projekt und im Vergleich mit weiteren Standards wie dem STTS und TIGER zu klären, ob die Anwendung entsprechender Modelle auch dazu führt, dass die theoretischen Besonderheiten der unterschiedlichen Annotationsschemata auch nach der maschinellen Ableitung erhalten bleiben, womit bei der Korpusnutzung ein Korrektiv zur Abschwächung des Ostereiproblems gegeben wäre. In jedem Fall stellen entsprechende Ergänzungen aber einen Schritt zur Erhöhung von Usability und Vergleichbarkeit dar. Sie bewirken einen Ausgleich zwischen projektinterner Forderung nach Innovation im Rahmen von Forschungsinteressen und dem Anspruch der Forschungsgemeinschaft auf eine gut in die digitale Forschungsinfrastruktur eingebunde Ressource.

Literatur

Ágel, Vilmos (2000): Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 2., vollst. neu bearb. u. erw. Aufl. 2. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 2.2). Berlin/New York: De Gruyter, S. 1855–1903.

Ágel, Vilmos (2017): Grammatische Textanalyse. Textglieder, Satzglieder, Wortgruppenglieder. Berlin/Boston: De Gruyter.

Ágel, Vilmos (2019): Grammatische Textanalyse (GTA) – eine deszendente Syntax des Deutschen. In: Eichinger, Ludwig M./Plewnia, Albrecht (Hg.): Neues vom heutigen Deutsch. Empirisch – methodisch – theoretisch. (= Jahrbuch des Instituts für Deutsche Sprache 2018). Berlin/Boston: De Gruyter, S. 265–291.

Ágel, Vilmos (2022): Richtlinien für die Textvorbereitung im DFG-Projekt „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2021/12/richtlinien-fuer-die-textvorbereitung_gieskane.pdf (Stand: 2.8.2022).

Ágel, Vilmos/Hennig, Mathilde (2022): Annotationshandbuch des DFG-Projekts „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2022/01/annotationsrichtlinien.pdf (Stand: 2.8.2022).

Ágel, Vilmos/Höllein, Dagobert (2021): Satzbaupläne als Zeichen: die semantischen Rollen des Deutschen in Theorie und Praxis. In: Binanzer, Anja/Gamper, Jana/Wecker, Verena (Hg.): Prototypen – Schemata – Konstruktionen. Untersuchungen zur deutschen Morphologie und Syntax. (= Reihe Germanistische Linguistik 325). Berlin/Boston: De Gruyter, S. 125–251.

Albert, Stefanie/Anderssen, Jan/Bader, Regine/Becker, Stephanie/Bracht, Tobias/Brants, Sabine/Brants, Thorsten/Demberg, Vera/Dipper, Stefanie/Eisenberg, Stephan/Hansen, Silvia/Hirschmann, Hagen/Janitzek, Juliane/Kirstein, Carolin/Langner, Robert/Michelbacher, Lukas/Plaehn, Oliver/Preis, Cordula/Pußel, Marcus/Rower, Marco/Schrader, Bettina/Schwartz, Anne/Smith, George/Uszkoreit, Hans (2003): TIGER Annotationsschema. https://www.ims.uni-stuttgart.de/documents/ressourcen/korpora/tiger-corpus/annotation/tiger_scheme-syntax.pdf (Stand: 2.8.2022).

Barteld, Fabian/Ihden, Sarah/Dreesen, Katharina/Schröder, Ingrid (2018): HiNTS: A tagset for Middle Low German. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara. Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, Miyazaki, Japan. Paris: European Language Resources Association (ELRA), S. 3940–3945. http://www.lrec-conf.org/proceedings/lrec2018/summaries/870.html (Stand: 2.8.2022).

Brants, Thorsten (2000): Inter-annotator agreement for a German newspaper corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), May 31–June 2, 2000, Athens, Greece. Paris: European Language Resources Association (ELRA), S. 165–172.

DeReKo (2022): Das Deutsche Referenzkorpus – DeReKo. https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (Stand: 2.8.2022).

DFG-Fachkollegium 104 „Sprachwissenschaften“ (2019): Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (Stand: 2.8..2022).

Dipper, Stefanie/Kwekkeboom, Sarah (2018): Historische Linguistik 2.0. Aufbau und Nutzungsmöglichkeiten der historischen Referenzkorpora des Deutschen. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Korpuslinguistik. (= Germanistische Sprachwissenschaft um 2020 5). Berlin/Boston: De Gruyter, S. 95–124.

Dipper, Stefanie/Donhauser, Karin/Klein, Thomas/Linde, Sonja/Müller, Stefan/Wegera, Klaus-Peter (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: Journal for Language Technology and Computational Linguistics 28, 1, S. 85–137. https://jlcl.org/content/2-allissues/10-Heft1-2013/5Dipper.pdf (Stand: 2.8..2022).

DTA (2022): Deutsches Textarchiv. https://www.deutschestextarchiv.de/doku/ueberblick (Stand: 2.8.2022).

Eisenberg, Peter (2020): Grundriss der deutschen Grammatik. Bd. 2: Der Satz. 5., aktual. u. überarb. Auflage. Unter Mitarbeit von Rolf Schöneich. Stuttgart/Weimar: Metzler.

Eisenberg, Peter/Lezius, Wolfgang/Smith, George (2005): Die Grammatik des TIGER-Korpus. In: Schwitalla, Johannes/Wegstein, Werner (Hg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv: Würzburger Kolloquium 2003. Tübingen: Niemeyer, S. 81–87.

Elspaß, Stephan (2012): Wohin steuern Korpora die Historische Sprachwissenschaft? Überlegungen am Beispiel des ‚Neuhochdeutschen‘. In: Maitz, Péter (Hg.): Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. (= Studia Linguistica Germanica 110). Berlin/Boston: De Gruyter, S. 201–225.

Emmrich, Volker (i. Vorb.): GiesKaNe – Syntactic basic structures of New High German: natural language processing in the process of annotation.

Emmrich, Volker/Hennig, Mathilde (i. Dr.): Das Fokusglied. Ein Vorschlag zur satz- und wortgruppengrammatischen Funktion der Grad- bzw. Fokuspartikel. Unter Mitarbeit von Nilüfer Cakmak und Philipp Meisner. In: Deutsche Sprache 51.

Engelberg, Stefan/Meliss, Meike/Proost, Kristel/Winkler, Edeltraut (Hg.) (2015): Argumentstruktur zwischen Valenz und Konstruktion. (= Studien zur Deutschen Sprache 68). Tübingen: Narr.

Geyken, Alexander/Boenig, Matthias/Haaf, Susanne/Jurish, Bryan/Thomas, Christian/Wiegand, Frank (2018): Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hg.): Digitale Infrastrukturen für die germanistische Forschung. (= Germanistische Sprachwissenschaft um 2020 6). Berlin/Boston: De Gruyter, S. 219–248.

Gries, Stefan/Berez, Andrea (2017): Linguistic annotation in/for corpus linguistics. In: Ide, Nancy/Pustejovsky, James (Hg.): Handbook of linguistic annotation. Dordrecht: Springer, S. 379–409.

Höllein, Dagobert/Lotzow, Stephanie (2019): Tagset des DFG-Projekts „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2022/01/tagset_gieskane-1.pdf (Stand: 2.8.2022).

Jacobs, Joachim (2008): Wozu Konstruktionen? In: Linguistische Berichte 213, S. 3–44.

Kajuk (2009): Kasseler Junktionskorpus. https://www.uni-giessen.de/fbz/fb05/germanistik/absprache/sprachtheorie/kajuk (Stand: 2.8.2022).

Lafferty, John/McCallum, Andrew/Pereira, Fernando (2001): Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Brodley, Carla E./Pohoreckyj Danyluk, Andrea (Hg.): Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). San Francisco: Morgan Kaufmann, S. 282–289.

Leech, Geoffrey (2004): Developing linguistic corpora: a guide to good practice. https://users.ox.ac.uk/~martinw/dlc/chapter2.htm (Stand: 2.8.2022).

Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik: eine Einführung. 3., überarb. u. erw. Auflage. (= Narr Studienbücher). Tübingen: Narr.

Odebrecht, Carolin (2017): Metadaten und Standardisierung von historischen Korpora. Vortrag auf der Tagung „Referenzkorpora des Deutschen: Konzepte, Methoden, Perspektiven" Rauischholzhausen.

Pustejovsky, James/Stubbs, Amber (2012): Natural language annotation for machine learning: a guide to corpus-building for applications. Beijing u. a.: O'Reilly.

Reichmann, Oskar (1988): Zur Vertikalisierung des Varietätenspektrums in der jüngeren Sprachgeschichte des Deutschen. Unter Mitwirkung von Christiane Burgi, Martin Kaufhold und Claudia Schäfer. In: Munske, Horst Haider/Polenz, Peter von/Reichmann, Oskar/Hildebrandt, Reiner (Hg.): Deutscher Wortschatz. Lexikologische Studien. Ludwig Erich Schmitt zum 80. Geburtstag von seinen Marburger Schülern. Berlin/New York: De Gruyter, S. 151–180.

Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): A gold standard corpus of Early Modern German. In: Proceedings of the Fifth Law Workshop, LAW V, Portland, Oregon, 23–24 June 2011. Stroudsburg: Association for Computational Linguistics, S. 124–128.

Schiller, Arne/Teufel, Simone/Stöckert, Christian/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart/Tübingen: Universität Stuttgart/Universität Tübingen. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf (Stand: 2.8.2022).

Wegera, Klaus-Peter (2013): Language data exploitation: design and analysis of historical language corpora. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard. J. (Hg.): New methods in historical corpora. (= Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache 3). Tübingen: Narr, S. 55–73.

Welke, Klaus (2011): Valenzgrammatik des Deutschen. Eine Einführung. (= De Gruyter Studium). Berlin/New York: De Gruyter.

Westpfahl, Swantje/Schmidt, Thomas/Jonietz, Jasmin/Borlinghaus, Anton (2017): STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf (Stand: 2.8.2022).