GraphVar – Korpusaufbau und Annotation
Version 1.0

Kristian Berg Jonas Romstadt Cedrek Neitzert

Institut für Germanistik, Vergleichende Literatur- und Kulturwissenschaft, Rheinische Friedrich-Wilhelms-Universität Bonn

7. September 2021

Zusammenfassung

Dieses Dokument beschreibt den Aufbau und die Annotation des Korpus GraphVar, das aus über 1.600 Abiturklausuren besteht, die zwischen 1923 und 2018 geschrieben wurden.

1 Übersicht

GraphVar ist ein Korpus aus über 1.600 Abiturarbeiten, die zwischen 1923 und 2018 an einem niedersächsischen Gymnasium geschrieben wurden. Der Aufbau dieses Korpus wurde von 2016 bis 2019 von der Deutschen Forschungsgemeinschaft gefördert. Das Hauptinteresse beim Aufbau bestand in der Beobachtung und Beschreibung graphematischer Variation und ihrer Entwicklung über die Zeit; der Name leitet sich aus diesem Interesse ab. Leitend war die Frage, was Schreiberinnen und Schreiber eigentlich tatsächlich machen bzw. gemacht haben – und zwar unbeeinflusst von technischen Hilfsmitteln oder Schluss- und Endredakteuren, aber unter vergleichbaren Bedingungen. Neben schriftlinguistischen Fragestellungen ist das Korpus prinzipiell auch für syntaktische, morphologische und lexikalische Fragestellungen geeignet; auch didaktische Untersuchungen sind möglich, genau wie kulturwissenschaftliche. Das Korpus bietet ein Fenster auf den unverfälschten Schreibgebrauch von Abiturientinnen und Abiturienten im Laufe der Zeit.

Der Zugang zum Korpus kann formlos per Mail (graphvar@uni-bonn.de) beantragt werden; dabei muss das wissenschaftliche Interesse nachgewiesen werden. Schildern Sie dazu bitte lediglich kurz Ihren fachlichen Hintergrund (Fachrichtung/Position, z. B. Studentin der Germanistik, Doktorandin der Fachdidaktik Deutsch etc.) und Ihre Fragestellung.

2 Datengrundlage

Das GraphVar-Korpus enthält digitalisierte Texte von Abiturklausuren, die an einem niedersächsischen Gymnasium verfasst wurden. Die 1.617 Texte wurden zwischen 1923 und 2018 verfasst. Sie lassen sich den Fächern Deutsch, Biologie und Geschichte zuordnen. Vor 1950 wurden alle verfügbaren Abiturklausuren digitalisiert und in das Gesamtkorpus aufgenommen, nach 1950 wurden die Texte jeweils (meist) im Fünfjahresabstand ausgewählt.

Zum aktuellen Zeitpunkt sind 1.398 Abiturklausuren mit 2.424.275 Tokens (inkl. Satzzeichen) abrufbar. Das entspricht allen digitalisierten Arbeiten, die zwischen 1948 und 2018 verfasst wurden. Die vorliegenden Abiturarbeiten, die vor dem Ende des 2. Weltkrieges (also zwischen 1923–1947) verfasst wurden, werden aktuell eingepflegt.

Abbildung 1 zeigt die Verteilung der verfügbaren Texte auf die Jahrgänge und Fächer; Abbildung 2 zeigt die Verteilung der Texte auf das Geschlecht der Autorinnen und Autoren.

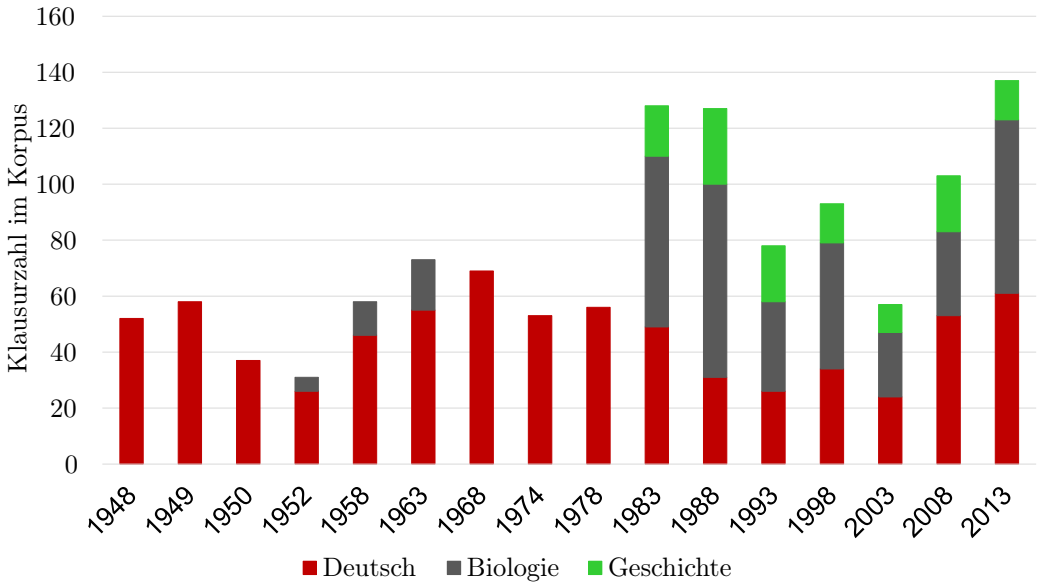

Abbildung 1: Anzahl der Klausuren im Korpus über die Zeit, geordnet nach Fächern.

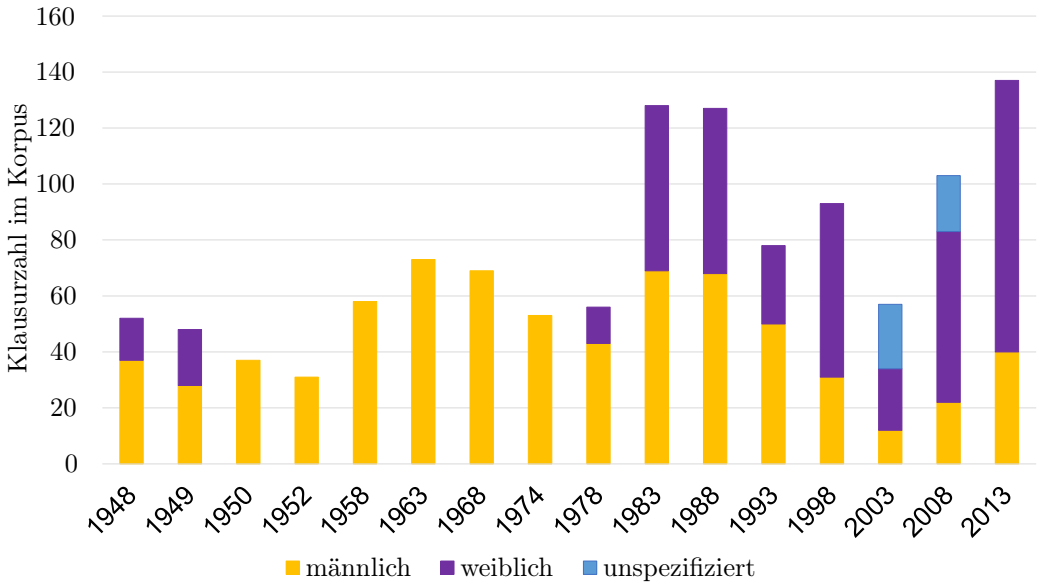

Abbildung 2: Anzahl der Klausuren im Korpus über die Zeit, geordnet nach Geschlecht.

Jahrgang	Deutsch		Biologie		Geschichte	
1948	37 M	15 W	–	–	–	–
1949	38 M	20 W	–	–	–	–
1950	37 M	0 W	–	–	–	–
1952	26 M	0 W	5 M	0 W	–	–
1958	46 M	0 W	12 M	0 W	–	–
1963	55 M	0 W	18 M	0 W	–	–
1968	69 M	0 W	–	–	–	–
1974	53 M	0 W	–	–	–	–
1978	43 M	13 W	–	–	–	–
1983	23 M	26 W	36 M	25 W	10 M	8 W
1988	20 M	11 W	33 M	36 W	15 M	12 W
1993	13 M	13 W	21 M	11 W	16 M	4 W
1998	5 M	29 W	17 M	28 W	9 M	5 W
2003	12 M	12 W	23 unspezifiziert		0 M	10 W
2008	10 M	43 W	12 M	18 W	20 unspezifiziert	
2013	13 M	48 W	21 M	41 W	6 M	8 W
2018	26 M	57 W	30 M	32 W	15 M	16 W

Tabelle 1: Anzahl der Arbeiten pro Jahrgang im Korpus, differenziert nach Geschlecht der Autorinnen und Autoren.

Die Dateinamen der einzelnen Arbeiten geben jeweils Auskunft über die jeweils spezifische Konfiguration der Metadaten pro Arbeit. Die Arbeit mit der Kennung 2013_DE_LK1_13_M_07P wurde etwa im Jahr 2013, im Fach Deutsch, im Leistungskurs mit der lf. Nr. 1 (LK1) durch den Schüler mit der lf. Nr. 13 verfasst. Es handelt sich dabei um einen männlichen Autoren (M), dessen Arbeit mit 07 Punkten (07P), also knapp befriedigend, bewertet wurde.

Die Biologie-Klausuren aus dem Jahrgang 2003 sowie die Geschichtsklausuren des Jahrgangs 2008 sind hinsichtlich des Geschlechts der Autorinnen und Autoren unspezifiziert. Mit Ausnahme des Geschlechts liegen keine weiteren Metadaten zu den Schreiberinnen und Schreibern vor, sodass nicht ausgeschlossen werden kann, dass Einzelne auch mit mehr als einer Klausur im Korpus vertreten sind (z. B. als Teil eines Deutsch-Grundkurses und eines Biologie-Leistungskurses).

Auch die Verteilung der Fächer ist nicht einheitlich. Nur Deutsch-Klausuren liegen durchgängig vor. Das ist letztlich ein Spiegel der historischen Entwicklung von Abiturklausuren allgemein. Bis in die 1970er-Jahre war vorgeschrieben, dass alle Schülerinnen und Schüler ihr Abitur in den Fächern Deutsch, Mathematik und Latein schreiben mussten (bzw. im neusprachlichen Zweig zusätzlich in Französisch, im altsprachlichen zusätzlich in Griechisch).

1972 wurde mit der Bonner Vereinbarung zur Neugestaltung der Oberstufe die gymnasiale Oberstufe schließlich reformiert und das bis heute bekannte Kurssystem wurde eingeführt (vgl. Kultusministerkonferenz 2021).[1] Damit ergaben sich auch Änderungen in den Vorgaben zur Abiturprüfung.

[1] Für eine genauere Rekonstruktion der institutionellen Einflussfaktoren und Beschlüsse der Kultusministerkonferenz

Deutsch-Klausuren liegen also durchgängig – sowohl vor als auch nach der Oberstufenreform – vor. Mathematik- und Latein-Klausuren, für die das ebenso gilt, enthalten entweder nur verhältnismäßig wenig Fließtext oder sind fremdsprachlich.

Ab 1952 wurden zudem auch Biologie-Klausuren digitalisiert. Biologie ist unter den Naturwissenschaften dasjenige Fach, dessen Klausuren am meisten Fließtext enthält. Ab 1983 werden auch Geschichts-Klausuren mitberücksichtigt, weil sie in einer größeren Quantität vorliegen als die anderer Gesellschaftswissenschaften.

Betrachtet man die Texte aus einer Makroperspektive, so ist zunächst feststellbar, dass sie immer länger werden, also immer mehr (graphematische) Wörter enthalten, wie Abbildung 3 zeigt.

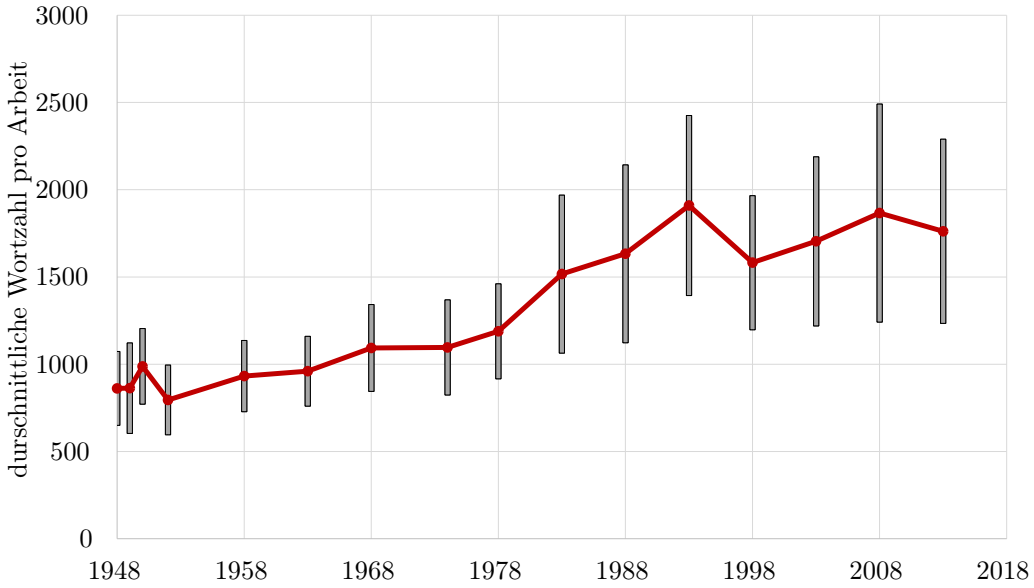

Abbildung 3: Durchschnittliche Wortanzahl pro Arbeit über die Zeit. Die Fehlerbalken entsprechen der Standardabweichung um den Mittelwert der Wortanzahl pro Arbeit.

Alle hier durchsuchbaren Texte sind Abiturklausuren. Diese grundlegende Feststellung hat methodische Vorteile, weil dieser Status nahelegt, dass einige externe Einflussfaktoren, die auf individuelle Schreibprozesse wirken können (vgl. Wrobel 2014), bei allen hier beschriebenen Texten vergleichbar sind. Die Schülerinnen und Schüler sind (vermutlich) durchgängig darauf bedacht, ein orthographisch korrektes Schreibprodukt für die Lehrkraft zu verfassen. Andere externe Faktoren, wie z. B. die für den konkreten Schreibprozess zur Verfügung stehende Zeit, haben sich im Laufe der Zeit verändert.

Neben diesen Rahmenvorgaben verändern sich auch die Aufgabenstellungen über die Zeit. Bis in die 1970er-Jahre hinein, müssen die Schülerinnen und Schüler sogenannte Besinnungsaufsätze verfassen (vgl. Ludwig & Merchert 1987). Darin sollen die Schreibenden über allgemeine Wertfragen reflektieren und so ihre Gesinnung begründet darstellen. Das ist nicht vergleichbar mit den heute an die Schülerinnen und Schüler herangetragenen Aufgabenstellungen, die sich gemäß aktueller Vorgaben der Kultusministerkonferenz grob in die Bereiche analysieren, interpretieren, erörtern und informieren/argumentieren unterteilen lassen (vgl. Steets 2014).

Und noch ein weiterer Aspekt wirkt sich auf die historische Vergleichbarkeit der Texte aus. In der pädagogischen Forschung ist häufig von einer Bildungsexpansion die Rede. Damit bezeichnet man den

zur Abiturprüfung nach dem 2. Weltkrieg siehe Wolter (2016), Böllig (2019).

Umstand, dass sich das Gymnasium von einer eher elitären hin zur seit den 1990er-Jahren meistbesuchten Schulform in Deutschland entwickelt hat (vgl. Becker 2000). Die Zusammensetzung der einzelnen Abiturjahrgänge schwankt also. Auch hier spiegelt sich folglich eine gesamtgesellschaftliche Entwicklung in den vorliegenden Daten. Einen Eindruck von diesem Prozess kann gewonnen werden, indem man ihn quantitativ erfasst. In erster Näherung kann etwa die Zahl derjenigen, die in einem spezifischen Jahr die Abiturprüfung ablegen als Anteil der Kohorte an Jugendlichen verstanden werden, die genau 18 Jahre zuvor an diesem zu definierenden Ort geboren wurden. Das ist dann notwendigerweise eine grobe Annäherung, weil Migrationsbewegungen jeglicher Natur unberücksichtigt bleiben. Führt man diese Berechnung durch, ergibt sich für das Bundesland Niedersachsen sowie für den Landkreis, in dem die hier untersuchten Abiturarbeiten entstanden sind, das Bild in Abbildung 4.

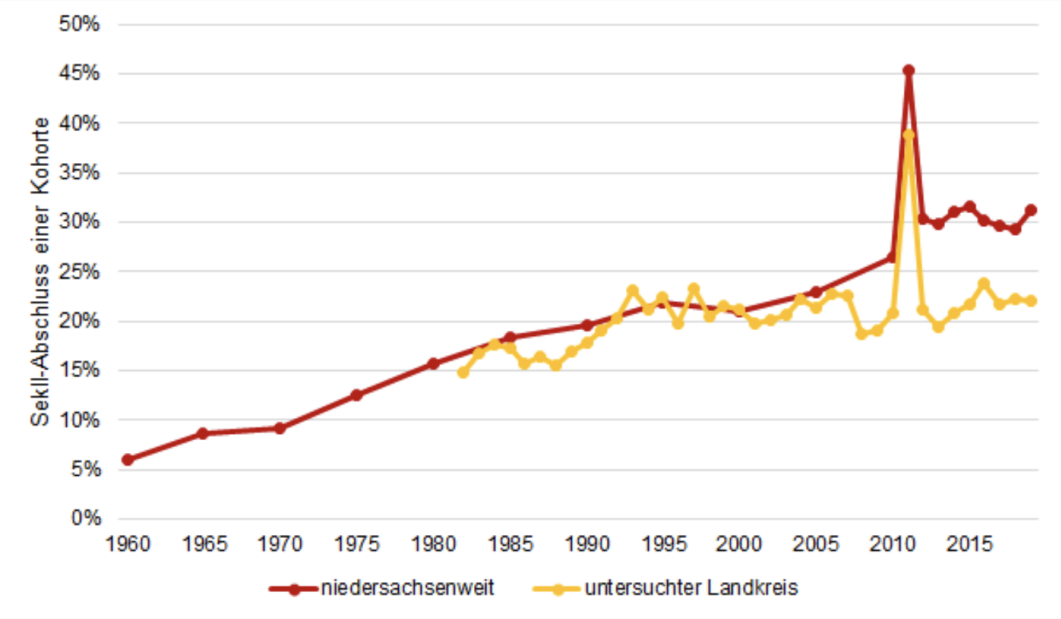

Abbildung 4: Anteil der pro Jahrgang im betreffenden Landkreis geborenen 18jährigen Personen, die ihre Schullaufbahn mit dem Abitur beenden, nach Berg & Romstadt 2021: 220, Datengrundlage: Niedersächsisches Landesamt für Statistik.

Niedersachsenweit steigt der Anteil derjenigen 18-jährigen, die in einem bestimmten Jahr eine Abiturprüfung[2] absolvieren von knapp über 5 % in den 1960er-Jahren über ein Fünftel der Kohorte in den 1990er-Jahren auf mittlerweile über 30 %. Der ungewöhnlich hohe Anteil an Abiturientinnen und Abiturienten im Jahr 2011 kann mit der Umstellung vom acht- zum neunjährigen Gymnasium begründet werden, sodass 2011 ein doppelter Jahrgang zum Abitur zugelassen wurde. Der Anteil an Abiturientinnen und Abiturienten an allen Personen, die 18 Jahre zuvor in Niedersachsen geboren wurden, steigt also merklich an. Für den Landkreis, in dem die hier beschriebenen Abiturarbeiten verfasst wurden, gilt das zwar auch, allerdings nicht im gleichen Maße. Seit den 1990er-Jahren liegt der Anteil an Abiturientinnen und Abiturienten an ihrer Kohorte – mit Ausnahme von 2011 – stets zwischen 19 % und 24 %; ein größerer Anstieg ist nicht zu beobachten. Das erhöht grundsätzlich die Vergleichbarkeit der hier dargestellten Daten – zumindest seit 1990.

[2] Für diese Auswertung wurden die Prüfungen zur Allgemeinen Hochschulreife („Abitur“) sowie zur Fachhochschulreife („Fachabitur“) kumuliert betrachtet.

3 Metadaten

In den Metadaten sind folgende Informationen festgehalten:

Metadaten	Kategorien
Jahr	1948, 1949, 1950, 1952, 1958, 1963, 1968, 1974, 1978, 1983, 1988, 1993, 1998, 2003, 2008, 2013, 2018
Fach	Biologie, Deutsch, Geschichte
Geschlecht	M (‚männlich'), W (‚weiblich'), X (‚keine Information')
Punkte	00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15

Tabelle 2: Metadaten im Korpus GraphVar.

Die Metadaten lassen sich verwenden, um im Abfragesystem ANNIS zu filtern (um bspw. nur diejenigen Texte zu finden, die 1998 im Fach Geschichte geschrieben wurden und mit 11 Punkten bewertet wurden); dazu in Abschnitt 5 mehr. Die Metadaten sind außerdem in den xml-Dateien des EXMARaLDa-Eports eingebettet (s. Abschnitt 4).

4 Workflow und Annotationsebenen

Neben den Metadaten gibt es für jede Klausure Annotationsebenen, mit denen die Texte in ganz unterschiedlicher Hinsicht beschrieben werden. Mit Digitalisaten der Klausuren als Vorlage wurden die nativen Texte der Klausuren zunächst transkribiert. Diese Ebene ist für alle weiteren Annotationen zentral; wir bezeichnen sie als `IST`-Ebene, weil hier die Texte mit all ihren Fehlern wiedergegeben sind. Die Transkription erfolgt allerdings nicht zeichengenau in Isolation – dann sind die Buchstaben ‹n› und ‹u› häufig nur schwer zu unterscheiden –, sondern im jeweiligen morphologischen und syntaktischen Kontext. So ist sichergestellt, dass die betreffenden Buchstaben und Wörter als sprachliche Einheiten wahrgenommen und adäquat transkribiert werden (vgl. Eisenberg 2020: 28 für Argumente gegen eine allzu „konkretistische" Reduktion).

Anschließend folgt auf einer zweiten Ebene die Normalisierung der sprachlichen Einheiten (`NORMAL`-Ebene). Der Zweck dieser Ebene ist es, jeder Einheit eine zeitlich invariante Form zuzuweisen. Die Konjunktion ‹dass› bspw. wurde vor 1996 regelmäßig mit finalem ‹ß› geschrieben. Über die `NORMAL`-Ebene sind beide Formen zu finden. Die Regel besagt, dass die seit 2006 gültigen Schreibungen als Zielform dienen. Sind mehrere Varianten freigegeben, folgt die Normierung den Empfehlungen im Rechtschreibduden.

Gleichzeitig mit der Normalisierung findet auf weiteren Ebenen die Kennzeichnung von Zitaten, Überschriften und Zeilentrennungen statt. Zitate und Überschriften sind als eigene Ebenen mittels binärer 1 oder 0 gekennzeichnet (Ebenen `ZITAT` und `ÜBERSCHRIFT`). Ein Wort, das sich im Originaltext über zwei Zeilen erstreckt, wurde durch einen Bindestrich an der entsprechenden Trennstelle auf der Ebene der `ZEILENTRENNUNG` gekennzeichnet (1).

(1) Versöhnungs-leistungen

Sind sprachliche Einheiten unleserlich, so dass keine zweifelsfreie Transkription gewährleistet werden kann, wird dies auf der Ebene `UNEINDEUTIG` mit binär 1 gekennzeichnet (andernfalls – und das ist die Regel – ist diese Ebene mit 0 gefüllt). Zuweilen kommentieren Annotatoren diese Ebene oder

anderweitige auffällige Umstände auf der Ebene `Kommentar`, wie z. B. der Fall, wenn ein Satz syntaktisch unvollständig ist und die Bereinigung auf der `NORMAL`-Ebene zu spekulativ geraten würde. Tabelle 2 fasst die oben diskutierten Annotationsebenen zusammen.

Ebene	Beispiel	Erläuterung
`IST`	`IST` = "Opterrituale"	Unveränderte Transkription der Worteinheit
`NORMAL`	`NORMAL` = "Opferrituale"	Normierte und korrigierte `IST`-Ebene bei Vorliegen eines Fehlers
`ÜBERSCHRIFT`	`ÜBERSCHRIFT` = "1"	Kennzeichnet, ob die Worteinheit Teil einer Überschrift ist; hier zutreffend; binäre Ausprägung
`ZITAT`	`ZITAT` = "1"	Zeigt an, ob die Worteinheit Teil eines Zitates ist; hier zutreffend; binäre Ausprägung
`ZEILENTRENNUNG`	`ZEILENTRENNUNG` = "Opterrituale"	Zeigt an, ob und wo ein Wort einen Zeilenumbruch erlebt hat
`UNEINDEUTIG`	`UNEINDEUTIG` = "1"	Markiert, ob eine Worteinheit unleserlich ist; hier zutreffend; binäre Ausprägung
`Kommentar`	`Kommentar` = "unvollständig"	Kommentarebene der Annotatoren

Tabelle 3: Annotationsebenen mit Beispielen und Erläuterungen.

Eine zweite Gruppe von Annotationsebenen bezieht sich auf Normen und Fehler. Fehler sind zeitabhängig, das muss mitbedacht werden. Was gestern falsch war, kann heute richtig sein, und umgekehrt. Das heißt, dass ausschließlich die orthografischen Regeln relevant sind, die zur Zeit der jeweiligen Abiturprüfung galten. Bis 1996 war das die jeweils aktuelle Auflage des Rechtschreib-Dudens; danach die Veröffentlichungen der Zwischenstaatlichen Kommission für deutsche Rechtschreibung bzw. des Rats für deutsche Rechtschreibung. Bis 1996 war bspw. die Subjunktion ‹daß› nur mit Eszett regelkonform, nach Inkrafttreten jedoch auch die Schreibung mit Doppel-s (‹dass›); nach Ablauf der zehnjährigen Übergangsfrist war letztere die einzige regelkonforme Form. Eine Suche nach dieser falsch geschriebenen Subjunktion führt auf der Ebene `Fehlerkategorie` vor und nach 1996 folglich zu unterschiedlichen Resultaten.

Die jeweils gültige Norm ist das orthographische Ziel der Texte. Sie wird auf der Ebene `IST_Ziel` annotiert. Diese Ebene ist gewissermaßen eine orthographisch ‚perfekte' Form des Textes. Verstößt eine Schreibung auf der `IST`-Ebene gegen die damals gültige Orthografie, wird dies auf der Ebene `Fehler` mit 1 markiert (ansonsten 0). Auf der Ebene `Fehlerkategorie` wird dann näher beschrieben, um welche Art Fehler es sich auf der `IST`-Ebene handelt. Semantik und Syntax – mit Ausnahme fehlender oder doppelter Wörter – bleiben unangetastet. Die Fehler verteilen sich auf sieben Kategorien und sind in den folgenden Beispielen unterstrichen.

Grundsätzlich kann ein und dieselbe sprachliche Einheit zwei Fehler enthalten. Diese sind in separaten Kategorien notiert, so dass im Korpus bis zu drei Fehlerannotationsspalten vorliegen. In Beispiel (2) fehlen sowohl das Genitivsuffix als auch die initiale Majuskel. Daher sind die zwei Fehlerkategorien Wortschreibung (`WORT`) und Flexionsfehler (`FLEX`) anzusetzen.

(2) Infolge eines zusammenschluss von ...

Unter Interpunktionsfehler (PKT) fallen fehlende oder falsch gesetzte Satzzeichen wie Kommata, (Doppel-)Punkte, Semikola, Gedankenstriche und Anführungszeichen. In Beispiel (3a) ist ein ‚überflüssiges' Komma vorhanden, während es in (3b) fehlt. Ähnliches gilt für den satzabschließenden Punkt hinter dem Zitationsort in (3c). In (3d) wiederum sind zitateinleitende Anführungszeichen vorhanden, jedoch keine abschließenden.

(3) a. In großen Krisen war man, auf andere Staaten angewiesen ...
b. Die Polizei war in der Lage den Täter zu finden.
c. ... ein mit „Mauern eingefriedeter Bezirk". (Z. 28-30) Die Hierarchie ...
d. ... „Ehrfurcht" und „höchste Verpflichtung (Z. 20) ...

Von der Kategorie der Interpunktionsfehler sind allerdings solche Fehler ausgeschlossen, die Bestandteil einer Abkürzung (4a) oder einer Ordinalzahl sind (4b). Hier handelt es sich um Fehler in der Wortschreibung (WORT). Zu dieser Kategorie zählen außerdem Wörter mit falschen Buchstaben (4c, 4d) oder fehlenden Buchstaben (4e). Onymische, pränominale Genitivattribute sind seit der Überarbeitung der Rechtschreibung 1996 erlaubt. Wenn ein Eigenname hervorgehoben werden soll, darf ein Apostroph Name und Flexiv abgrenzen. Bei andersgearteten Fällen nach 1996 und generell vor 1996 handelt es sich jedoch um Normverstöße, die in den Texten als Fehler markiert werden (4f).

(4) a. ... (vgl. Z 5) ...
b. Wie im 15 Jahrhundert ...
c. Reperaturkosten
d. Opterrituale
e. Christliche Symbole wurden mit den Göttern der Indigenen verschmolen.
f. Als Fazit zu Lutz Graf von Schwerin von Krosigk's Rundfunkansprache kann man sagen ...

Fehler in der Groß- und Kleinschreibung (GKS) beschreiben bspw. fälschlich großgeschriebene Verben (5a) oder kleingeschriebene Substantive (5b).

(5) a. ... ihre zerstörerische Haltung zu Rechtfertigen.
b. ... ob sie die Forderungen verstanden, hauptsache, es wurde vorgelesen.

Eine weitere Kategorie sind Fehler in der Getrennt- und Zusammenschreibung (GZS). Ob Wörter zusammen- (6a) oder getrenntgeschrieben (6b) werden, hängt bis 1996 von der im Rechtschreib-Duden angegebenen Regel bzw. dem Duden-Eintrag ab. Neben der Reform von 1996 stellt die Revision von 2004 eine Zäsur dar, denn innerhalb dieses Zeitraumes ist die Zusammenschreibung generell als Fehler zu werten.

(6) a. ... auch wenn dies mit Gewalt statt finden musste.
b. ... es muss im Gedächtnis haftenbleiben.

Fehler, die die Rektion und Kongruenz betreffen und sich innerhalb eines Satzes auflösen lassen, fallen unter die Kategorie Flexionsfehler (FLEX). Hierunter zählen alle flektierbaren Wortarten wie Artikel (7a), Verben (7b) und Substantive (7c).

(7) a. Diese Hyperbel wird durch einen Enjambement hervorgehoben.

b. Sein letzter Punkt ist, dass diese Modifikationen direkt an die nächste Generation vererbt wird.

c. ... und ihre Zuhörer sind die anwesenden Bundestagsabgeordnete.

Nicht korrigiert werden hierbei pronominale Wiederaufnahmen über Satzgrenzen hinaus (8).

(8) Das Mädchen trug ein rotes Käppchen. Sie/Es ging in den Wald.

In einigen Fällen treten fehlerhaft doppelte Wörter (`DOPPEL`) auf. Diese können direkt nebeneinander stehen (9a) oder auch darauf hinweisen, dass ein Satz ursprünglich abweichend konzipiert war (9b).

(9) a. Die spanische Herrschaftspraxis war von Gewalt und und Herrschsucht betrieben.

b. Somit ist bewiesen, dass die Arten untereinander Verwandt sind, jedoch sich durch Evolution viele DNA-Sequenzen durch verschiedene Faktoren sich verändert haben.

Z. T. fehlen bestimmte Wörter, in (10) z. B. das Hilfsverb *werden* am Satzende. In solchen Fällen enthält der Satz auf der `IST`-Ebene ein leeres Element, das auf der Ebene der Fehlerkategorie als `FEHL` (fehlendes Element) klassifiziert wird. Der Eingriff ist dabei so minimal wie möglich; denkbar sind ja (u. a.) auch *werden können*, *werden müssen*, *worden sein*.

(10) Darüber sollte diskutiert.

Tabelle 4 fasst die Annotation der Fehler zusammen:

Ebene	Beispiel	Erläuterung
`IST_Ziel`	`IST_Ziel` = "daß"	Gibt für jede laufende Wortform diejenige Form an, die zum Zeitpunkt des Schreibens orthographisch korrekt war
`Fehler`	`Fehler` = "1"	Zeigt an, ob die damals gültige `IST_Ziel`- und `IST`-Ebene divergieren; hier zutreffend; binäre Werte
`Fehlerkategorie`	`Fehlerkategorie` = "WORT"	Klassifikation der Fehler; im Beispiel Wortfehler. Außerdem möglich: `GKS`, `GZS`, `FLEX`,`PKT`, `DOPPEL`, `FEHL`

Tabelle 4: Übersicht der Ebenen der Fehlerannotation.

Eine dritte Gruppe von Annotationsebenen umfasst die Anreicherung der Texte mit lexikalischen, morphologischen und syntaktischen Informationen. Alle Klausuren haben ein Part-Of-Speech-Tagging (POS-Tagging) durchlaufen, sodass für jede laufende Wortform die Wortart und das jeweilige Lemma verfügbar sind (Ebenen `NORMALpos` und `NORMALlemma`). Im selben Arbeitsschritt wurden auch die Satzgrenzen markiert (Ebene `NORMALS`). Als Grundlage dafür diente das im Falko-Projekt entwickelte gleichnamige Annotationstool als Excel-AddIn (vgl. Reznicek 2012). Das POS-Tagging selbst erfolgte mithilfe des TreeTaggers (vgl. Schmid 1995). Die Beschreibung der Wortformen beruht auf dem Stuttgart-Tübingen-Tagset (STTS), das folgende 54 Klassen vorsieht (vgl. Schiller et al. 1999):

POS =	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	*um [zu leben],* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit,* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*
NN	Appellativa	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	substituierendes Relativpronomen	*[der Hund,] der*

Abbildung 5: STTS-Labels 1 (aus Schiller et al. 1999: 6).

POS =	Beschreibung	Beispiele
PRELAT	attribuierendes Relativpronomen Relativpronomen	*[der Mann ,] dessen [Hund]*
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe],* *wessen [Hut]*
PWAV	adverbiales Interrogativ– oder Relativpronomen	*warum, wo, wann,* *worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	"zu" vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten],* *zu [schnell]*
TRUNC	Kompositions–Erstglied	*An– [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	– [ ]()

Abbildung 6: STTS-Labels 2 (aus Schiller et al. 1999: 7).

Zusätzlich zur Annotation der Wortarten und Lemmata erlaubt das Falko-Tool die Markierung der Satzspannen und -grenzen. Da die automatische Annotation insbesondere der Wortarten relativ fehlerbehaftet ist, folgte eine manuelle Validierung und Korrektur durch linguistisch geschulte Hilfskräfte.[3]

In einem letzten Schritt wurden basale syntaktische Informationen annotiert, und zwar in Form von topologischen Positionen und Konstituentenklassen, die auf dem Feldermodell des Deutschen beruhen (für eine Einführung vgl. Wöllstein 2010). Technisch wurden die Sätze automatisch durch einen minimal modifizierten Berkeley Parser analysiert (vgl. Petrov et al. 2006), um eine vorläufige Felderstruktur zu generieren. Im Anschluss wurden die Tabellen mithilfe des Pepper-Konverters (vgl. Zipser & Romary 2010) in ein tsv-Format umgewandelt, das bei der internetbasierten Annotationsdistribution WebAnno Verwendung findet (vgl. Castilho et al. 2016). Dieses Tool ermöglicht die Annotation und Kuration von Projekten. Hier erfolgte eine händische Bearbeitung der Topologien der Texte und die Verbesserung

[3] Die Fehlerquote der durch den TreeTagger bestimmten Wortarten liegt bei $4{,}39 \pm 1{,}52$ Prozent und bei den Lemmata bei $4{,}87 \pm 3{,}03$ Prozent.

der automatisch erzielten Ergebnisse des Berkeley Parsers durch linguistisch trainierte Hilfskräfte.

Die Information über die topologische Position und die Gliederung in Konstituenten ist auf der Ebene `TopField` abrufbar. Das Vorfeld ist durch *VF* gekennzeichnet, das Mittelfeld durch *MF* und das Nachfeld durch *NF*. *LK* steht für die linke Satzklammer, in denen sich in Verbzweitsätzen das finite Verb und in Sätzen mit Verbletztstellung Subjunktionen wiederfinden. *VC* bezeichnet die rechte Satzklammer, in der in Verbletzt-Sätzen das finite Verb und in Verbzweitsätzen Prädikatsbestandteile zu finden sind. Konstituenten und Phrasen erhalten ihre Kategorie abhängig von ihrem lexikalischen Kopf: So werden Nominalphrasen mit *NX* und Präpositionalphrasen mit *PX* annotiert. Da Verbalphrasen in den Analysen als oberste Phrase konzeptualisiert werden (sofern sie denn ein Prädikat aufweisen), bezeichnet SIMPX sowohl die komplette Satzmatrix als auch die untergeordneten Verbzweit- und Verbletztsätze exklusive des satzschließenden Punktes, der als $. separiert auftritt. Liegt kein Prädikat vor, ist die komplette Satzmatrix nicht dem Tag SIMPX untergeordnet, da ein vollständiger Satz traditionell ein Verb benötigt. Eine tabellarische Auflistung aller verwendeten „Node labels“ findet sich in Telljohann et al. (2015: 25). Diese thematisiert u. a. das Vorgehen bei der Annotation von Topologien und Phrasen in der Tübinger Datenbank für geschriebenes Deutsch (TüBa-D/Z) und dient dem GraphVar-Projekt als Guideline-Dokumentation.

Node Labels	**Description**
Phrase Node Labels	
ADJX	adjectival phrase
ADVX	adverbial phrase
DP	determiner phrase (e.g. *gar keine*)
FX	foreign language phrase
NX	noun phrase
PX	prepositional phrase
VXFIN	finite verb phrase
VXINF	non-finite verb phrase
Topological Field Node Labels	
LV	resumptive construction (Linksversetzung)
C	complementizer field (C-Feld)
FKOORD	coordination consisting of conjuncts of fields
KOORD	field for coordinating particles
LK	left sentence bracket (Linke (Satz-)Klammer)
MF	middle field (Mittelfeld)
MFE	middle field between VCE and VC
NF	final field (Nachfeld)
PARORD	field for non-coordinating particles
VC	verb complex (Verbkomplex)
VCE	verb complex with the split finite verb of *Ersatzinfinitiv* constructions
VF	initial field (Vorfeld)
FKONJ	conjunct consisting of more than one field
Root Node Labels	
DM	discourse marker
P-SIMPX	paratactic construction of simplex clauses
R-SIMPX	relative clause
SIMPX	simplex clause

Abbildung 7: Node labels (aus Telljohann et al. 2015: 25).

Tabelle 5 fasst die lexikalischen, morphologischen und syntaktischen Annotationsebenen zusammen:

Ebene	Beispiel	Erläuterung
NORMALpos	NORMALpos = "NN"	Wortart der Worteinheit; hier: normales Nomen
NORMALS	NORMALS = "s56"	Verortung der Worteinheit in Satz Nummer 56
NORMALlemma	NORMALlemma = "Opferritual"	Das Lemma der Worteinheit der NORMAL-Ebene
TopField	TopField = "MF"	Zuordnung der Worteinheit innerhalb der Satzmatrix; hier: gehört zum Mittelfeld

Tabelle 5: Übersicht der Ebene der applikationsgestützten POS-Tags, Satzgrenzen, Lemmata sowie topologischer Information.

Neben den oben vorgestellten Annotationsebenen finden sich noch drei weitere, die vor allem für den internen Gebrauch konzipiert sind und die im Folgenden kurz besprochen werden sollen.

Die Ebene WebAnno ist binär und zeigt an, ob die Satztopologien durch Annotierende bereits validiert worden sind. Bei der Ausprägung 1 ist dies der Fall, bei der Ausprägung 0 steht eine Begutachtung noch aus.

Bei bestimmten Auswertungen ist die Variation der Textlänge, die sich über die Zeit beobachten lässt (s. oben), hinderlich: Nimmt die Fehlerzahl über die Zeit zu, liegt das u. U. schlicht an der steigenden Wortzahl der Klausuren. Um mit diesem Problem auf eine einfache Weise umzugehen, werden die laufenden Wörter pro Text gezählt (Ebene IST_index). So ist es möglich, bspw. jeweils nur in den ersten 1000 Wörtern jedes Textes zu suchen.

Die Ebene IST_NORMAL_DIFF zeigt Differenzen zwischen der IST- und der NORMAL-Ebene an (mit *XXX*). Das kann als Hinweis auf orthographische Fehler gesehen werden (ist allerdings weder eine hinreichende noch eine notwendige Bedingung, sondern nur ein zusätzliches Indiz).

Ebene	Beispiel	Erläuterung
IST_index	IST_index = "126"	Ordinalzahl der Worteinheit in der Klausur; hier Wort Nummer 126
WebAnno	WebAnno = "1"	WebAnno-Validierung; hier zutreffend; binäre Ausprägung
IST_NORMAL_DIFF	IST_NORMAL_DIFF = "XXX"	Unterschied zwischen IST- und NORMAL-Ebene

Tabelle 6: Übersicht der Annotationsebenen für den internen Gebrauch.

Abschließend durchliefen die Klausuren eine erneute Konvertierung, diesmal in das vom Korpussuchtool ANNIS verwendete Format (vgl. Krause & Zeldes 2016) und in ein exportierbares EXMARaLDA-Format (vgl. Schmidt & Wörner 2014). Die Verarbeitung via Pepper und WebAnno geschah serverseitig.

5 ANNIS

ANNIS ist das zentrale Werkzeug, mit dem auf das Korpus zugegriffen wird. Für einen kompletten Überblick sei hier auf das ANNIS-Handbuch verwiesen (Krause & Zeldes 2016); an dieser Stelle sollen lediglich die wichtigsten Funktionen angerissen werden. Es lassen sich Abfragen auf unterschiedlichen Ebenen miteinander verbinden. So können wir bspw. nach Interpunktionsfehlern suchen, die in Arbeiten von 2013 vorkommen. Eine solche Abfrage würde wie folgt aussehen:

(11) `Fehlerkategorie` = "PKT" & `meta::Jahr` = "2013"

In dieser Abfrage sind zwei einzelne Abfragen verknüpft. Die erste betrifft die Fehlerkategorie: Hier muss der Wert "PKT" (Interpunktionsfehler, s. S. 7–8) lauten. Das ist aber nicht alles, denn diese Abfrage in den annotierten Daten wird mit einer Abfrage der Metadaten verknüpft. Gesucht wird nur in Arbeiten, die aus dem Jahr 2013 stammen. Beide Abfragen werden mit dem kaufmännischen " und"verknüpft (&). Das ist der Standardfall, wenn Abfragen des sprachlichen Materials mit solchen der Metadaten verbunden werden – in anderen Fällen stehen andere Operatoren zur Verfügung. Es lassen sich beispielsweise zwei Abfragen auf dieselbe Einheit beziehen; das geschieht mit dem Identitätsoperator "_=_":

(12) `Fehlerkategorie` = "PKT" _=_ `IST` = "," & `meta::Jahr` = "2013"

Gesucht wird hier ein Element, dessen Fehlerkategorie mit "PKT" annotiert ist und das auf der `IST`-Ebene ein Komma enthält (und das nach wie vor aus einer Arbeit von 2013 stammt). Auf diese Weise lassen sich also zu viel gesetzte Kommas finden – tatsächlich vorkommende Kommas, die als Fehler klassifiziert sind.

Im Fall oben beziehen sich beide Abfragen auf dieselbe Einheit; häufig ist es aber notwendig, direkt aufeinander folgende Einheiten zu beschreiben. Dafür gibt es den Abstandsoperator ".", der die direkte Abfolge beschreibt (ohne Unterbrechung durch andere Wörter). Wir können beispielsweise fehlerhafte Kommas suchen, die vor einem ‹dass› auftreten:

(13) `Fehlerkategorie` = "PKT" _=_ `IST`=",". `IST`="dass" & `meta::Jahr` = "2013"

Es gibt noch eine Reihe weiterer nützlicher Operatoren, z. B. den Einschlussoperator (_i_), der prüft, ob eine Einheit Teil einer größeren Einheit ist (z. B. fehlende Kommas in Relativsätzen) oder die Links- bzw. Rechtsalignierung von Einheiten zweier Ebenen (_l_ und _r_); damit lassen sich beispielsweise alle Sätze suchen, die mit ‹Und› beginnen.
Die Abfragebeispiele sind eigentlich Abkürzungen, das sei hier der Vollständigkeit halber erwähnt: Eine vollständige Abfrage besteht aus a) einzelnen Kriterien (auch „Nodes“, also Knoten) und b) deren expliziter Verknüpfung („Edge“, also Kante). Je komplexer die Abfragen werden, desto sinnvoller ist eine saubere Trennung. Beispiel (13) sieht in „Langform“ so aus:

(14) `Fehlerkategorie` = "PKT" &
`IST` = "," &
`IST` = "dass" &
#1 _=_ #2 &
#2 . #3 &
`meta::Jahr` = "2013"

In den ersten drei Zeilen werden die Kriterien aufgelistet und in den folgenden zwei Zeilen verknüpft. Dabei wird mit #[Zahl] auf die Kriterien verwiesen; #2 . #3 bedeutet bspw., dass das dritte Kriterium sich auf eine Einheit bezieht, die der Einheit des zweiten Kriteriums folgt. Empfehlenswert ist der Einsatz des sog. Query Builders, mit dem sich ohne großen Aufwand auch komplexe Abfragen formulieren lassen.

Ist eine Abfrage erfolgreich, lässt sie sich auch exportieren („More“ > „Export“) und weiterverarbeiten; das ist wie in allen korpuslinguistischen Untersuchungen unbedingt notwendig. Die ANNIS-Abfrage ist nur der erste Schritt, diese Daten müssen aber noch einmal geprüft und ggf. gefiltert werden. Mit dem GridExporter lassen sich bspw. mehrere Ebenen gleichzeitig exportieren (die relevanten Ebenen werden kommasepariert im Feld „Annotation Keys“ spezifiziert; die Nummerierung kann mit `numbers = false` im Feld „Parameter“ unterdrückt werden). Die so exportierten Daten können (ggfs. nach einer Bereinigung in einem Texteditor) z. B. in Excel weiterverarbeitet werden.

Abfragbar sind zurzeit alle Annotationsebenen und alle Metadaten aller verfügbaren Arbeiten. Noch nicht verfügbar sind die PDF-Dokumente, also die Scans der Arbeiten. Perspektivisch werden aber auch diese Informationen eingepflegt werden.

6 Gütekriterien/Inter-Annotator-Agreement

Den Transkriptionen und Annotationen liegen vorher festgelegte Standards und Klassifizierungsrichtlinien zugrunde. Über Applikationen gewonnene Daten wie POS-Tags, Lemmata und Satztopologien werden durch linguistisch ausgebildete Hilfskräfte validiert. Dabei berücksichtigen sie sowohl die Guidelines des STTS (vgl. Schiller et al. 1999) als auch die des „Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)“ (Telljohann et al. 2015). Obgleich alles daran gesetzt wird, reliable Ergebnisse zu schaffen, bestehen zwischen den Resultaten der Annotierenden Differenzen, da Übertragungsfehler entstehen und Einschätzungen von Zweifelsfällen unterschiedlich ausfallen können. Um den Grad der Reliabilität festzustellen, fand die Erhebung von Inter-Annotator-Agreements (IAA) statt. Als Berechnungswert diente Fleiss' Kappa, das sich bei der Transkription und Fehlerannotation zwischen drei Ratern auf 0,65 beläuft, was einer belastbaren Übereinstimmung entspricht. Bislang konnte jedoch kein IAA unter Berücksichtigung der Wortartenvalidierung und Topologiebestimmung durchgeführt werden.

Literatur

Becker, Rolf (2000): Bildungsexpansion und Bildungsbeteiligung. Oder: Warum immer mehr Schulpflichtige das Gymnasium besuchen. In: *Zeitschrift für Erziehungswissenschaft 3*, S. 447–480.

Berg, Kristian & Romstadt, Jonas (2021): Reifezeugnis – Das Komma in Abituraufsätzen von 1948 bis heute. In: Deutsche Akademie für Sprache und Dichtung/Union der deutschen Akademien der Wissenschaften (Hg.). *Die Sprache in den Schulen – Eine Sprache im Werden. Dritter Bericht zur Lage der deutschen Sprache*. Berlin: Erich Schmidt, S. 205–236.

Bölling, Rainer (2019): Die Reformen des Abiturs in Deutschland: eine Untersuchung zu Prüfungsfächern und Hochschulzulassung. In: *Global Education 48.8*, S. 100–115. http://www.rboelling.

de/download/B%C3%B6lling$_{A}$bitur%20in%20Deutschland$_{2}$019.pdf, 01.09.2021

Castilho, Eckart de et al. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In: *Proceedings of the LT4DH workshop at COLING 2016*. Osaka, Japan, S. 76-84. https://webanno.github.io/webanno/publications/W16-4011.pdf, 01.09.2021. WebAnno verfügbar unter: https://webanno.github.io/webanno/, 01.09.2021.

Eisenberg, Peter (2020): *Der Satz. Grundriss der deutschen Grammatik. 5., aktualisierte und überarbeitete Auflage*. Stuttgart: Metzler.

Krause, Thomas & Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities 2016.31*, S. 118-139. http://dsh.oxfordjournals.org/content/31/1/118, 01.09.2021. ANNIS3 verfügbar unter: https://corpus-tools.org/annis/, 01.09.2021.

Kultusministerkonferenz (2021): *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung. Beschluss der Kultusministerkonferenz vom 07.07.1972 i. d. F. vom 18.02.2021*. https://www.kmk.org/fileadmin/veroeffentlichungen$_{b}$eschluesse/1972/1972$_{0}$7$_{0}$7-VB-gymnasiale-Oberstufe-Abiturpruefung.pdf, 01.09.2021.

Ludwig, Otto & Merchert, Eckehart (1987): Fritz Rahn und der Besinnungsaufsatz. Zur Herkunft und Bedeutung einer Aufsatzform. In: *Praxis Deutsch 84*, S. 10–16.

Petrov, Slav et al. (2006): Learning Accurate, Compact, and Interpretable Tree Annotation. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, S. 433–440. https://aclanthology.org/P06-1055.pdf, 01.09.2021. Berkeley Parser verfügbar unter: https://github.com/slavpetrov/berkeleyparser, 01.09.2021.

Reznicek, Marc (2012): *Falko Annotation Excel-AddIn. Version 0.1.5. MS-O ce 2003–2003*. https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/marc/Falko$_{0}$.1.5.xla, 01.09.2021.

Schiller, Anne et al. (1999): *Guidelines für das Tagging deutscher Textcorpora [sic] mit STTS (Kleines und großes Tagset)*. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf, 01.09.2021.

Schmid, Helmut (1995): *TreeTagger - a part-of-speech tagger for many languages*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, 01.09.2021.

Schmidt, Thomas & Wörner, Kai (2014): „EXMARaLDA“. In: *Handbook on Corpus Phonology*, S. 402–419. http://ukcatalogue.oup.com/product/9780199571932.do, 01.09.2021.

Steets, Angelika (2014): Schreiben in der Sekundarstufe II. In: Feilke, Helmuth & Pohl, Thorsten (Hg.). *Schriftlicher Sprachgebrauch – Texte verfassen*. Baltmannsweiler: Schneider, S. 178–194.

Telljohann, Heike et al. (2015): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Tübingen. http://www.sfs.uni-tuebingen.de/fileadmin/user_upload/ascl/tuebadz-stylebook-1508.pdf, 01.09.2021.

Wöllstein, Angela (2010): *Topologisches Satzmodell*. Heidelberg: Winter.

Wolter, Andrä (2016): Gymnasium und Abitur als „Königsweg“ des Hochschulzugangs: Historische Entwicklungslinien und institutionelle Transformationen. In: Kramer, Jochen & Neumann, Marko & Trautwein, Ulrich (Hg.): *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre E ekte*. Wiesbaden: Springer, S. 1–28.

Wrobel, Arne (2014): Schreibkompetenz und Schreibprozess. In: Feilke, Helmuth & Pohl, Thorsten (Hg.). *Schriftlicher Sprachgebrauch – Texte verfassen*. Baltmannsweiler: Schneider, S. 85–100.

Zipser, Florian & Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards*, LREC 2010. La Valette, Malta. https://hal.inria.fr/inria-00527799/document, 01.09.2021. Pepper-Konverter verfügbar unter: https://corpus-tools.org/pepper/, 01.09.2021.