https://doi.org/10.21248/idsopen.16.2026.75

Laura Duve/Valeria Schick

Wandel im Pronomengebrauch vom Frühneuhochdeutschen zum frühen Neuhochdeutschen

Korpusaufbereitung und Annotation mit *INCEpTION* und Auswertung und Visualisierung mit *ANNIS*

Abstract This paper presents the workflow created for the corpus preparation and annotation by two projects of the DFG-funded research group *Practices of Personal Reference*. Both projects examine diachronic developments in the use of pronouns deployed for personal reference, with one project compiling a corpus of German dramas from the 17th–19th centuries and the other collecting a corpus of German medical texts from the 15th–17th centuries. Operating with the tools *INCEpTION* and *ANNIS*, the data will be annotated and analyzed following project-specific linguistic categories. The methodological approach will be illustrated by an exemplary study of so-called *Heische*-formulae which contain the impersonal pronoun *man* and contextualize various acts of request. The comparison of different text types, represented in the projects, allows to draw conclusions about functional and contextual differences in the pronoun usage across time and intends to show a way in which cross-project research aims can be realized successfully.

Keywords OCR, *INCEpTION*, *ANNIS*, historic corpora, annotation

1. Einleitung

In diesem Beitrag beschreiben wir die Erstellung und Aufbereitung zweier historischer Korpora von der Datensammlung bis hin zur Veröffentlichung der annotierten Daten. Die Korpora entstammen zwei historisch ausgerichteten Teilprojekten der DFG-geförderten Forschungsgruppe *Praktiken der Personenreferenz* (Projektnummer: 457855466).[1] Beide Projekte setzen sich mit Wandelprozessen im personenreferierenden Pronomengebrauch auseinander.

Das Teilprojekt II *Das Pronomen ‚man' in der Diachronie des Deutschen* (im Folgenden *man*-Projekt) untersucht die Entstehung und Verfestigung von Referenzspektren und Funktionen des impersonalen Pronomens *man* sowie deren Herausbildung in verschiedenen frühneuzeitlichen Gattungen als Produkt wiederkehrender kommunikativer Bedarfe. Für diesen Gattungsvergleich wird u. a. ein Korpus von Pesttraktaten und Bäderkunden erstellt. Bei beiden Gattungen handelt es sich um medizinisch-wissenschaftliche Texte, die sich sowohl an Laien als auch an medizinisches Personal richten. Erstere beschreiben Symptomatik und Behandlungsmöglichkeiten der Pest (vgl. Eckkrammer 2015, S. 38), während zweitere sich mit Heilbädern und ihrer Anwendung beschäftigen (vgl. Fürbeth 2012, S. 194). Da

1 Informationen zur Forschungsgruppe und den beteiligten Teilprojekten finden sich unter: https://forschungsgruppe-pronomen.de (Stand: 3.12.2025).

beide Gattungen Instruktionen beinhalten und diese Kontexte eine wichtige funktionale Domäne von *man* darstellen, eignen sie sich besonders gut für die Ziele des Projekts.

Das Teilprojekt IV *Personenreferierende Pronomen in Dramen – interaktionale und dramaturgische Funktionen sowie historischer Wandel von Barock über Aufklärung zu Sturm und Drang und Klassik* (im Folgenden Dramen-Projekt) analysiert aus einer historisch-interaktionalen Perspektive Veränderungen im Gebrauch von Personal-, Indefinit- und Demonstrativpronomen im Zeitraum vom 17.–19. Jahrhundert. Von Interesse ist dabei z. B., welche sprachlichen Muster die fokussierten Pronomengruppen in dieser Zeit hervorgebracht haben und welche interaktionalen und dramaturgischen Funktionen damit erfüllt wurden. Als Datengrundlage der Projektarbeit dienen Dramentexte (Komödien und Tragödien) von vier einschlägigen Autoren aus der Epochenspanne Barock bis Klassik.

Der Fokus dieses Beitrags liegt einerseits auf der Darstellung des in den Projekten erstellten Workflows und der verwendeten Tools. Dazu existieren insbesondere für historische Textgattungen bisher wenige umfassende Beschreibungen, da sich daraus spezifische Anforderungen an nahezu alle Schritte der Korpuserstellung und -aufbereitung ergeben. Andererseits möchten wir auch auf Fallstricke und datenspezifische Probleme hinweisen, die sich beim Verfolgen pragmatischer Fragestellungen in einem derartigen Rahmen ergeben können. Insgesamt zeigt unser Vorgehen einen probaten Weg auf, wie korpusbasierte Forschungsvorhaben mit historischen Daten projektübergreifend und in größeren Teams zielführend realisiert werden können. Die von uns beschriebenen Schritte und Tools können sowohl im Ganzen als auch in einzelnen Punkten adaptiert und an eigene Forschungsfragen und Daten angepasst werden.

Zum Abschluss wollen wir den Nutzen unseres Vorgehens für pragmatische Fragestellungen exemplarisch an sogenannten Heische-Formeln mit *man* illustrieren. Diese weisen die Grundstruktur *man + Verb in 3. Pers. Sg. Konj. I* auf und kontextualisieren verschiedene Formen von direktiven Handlungen (vgl. Duve/Schick eingereicht). Abgesehen von Beschreibungen in Grammatiken gibt es bisher wenige empirische Untersuchungen zu diesem Phänomen (vgl. z. B. Jäger 1971), wobei v. a. seine historische Entwicklung auf formaler und funktionaler Ebene noch wenig Beachtung gefunden hat (vgl. aber Duve/Schick eingereicht). Ausgehend von der genannten Untersuchung werden wir daher in Kapitel 6 eine Fallstudie zu diesem Thema vorstellen. Zunächst wird jedoch der Prozess der Datensammlung und -vorbereitung mithilfe verschiedener Datenbanken und automatischer Texterkennung beschrieben. In Kapitel 3 wird die Korpusanreicherung durch automatisches Tagging skizziert. Daraufhin wird auf das Annotationsvorgehen eingegangen, das projektübergreifend durch das Annotationstool *INCEpTION* (vgl. Klie et al. 2018) realisiert wird. Kapitel 5 erläutert die Auswertung beider Datensets in dem Korpusanalysetool *ANNIS* (vgl. Krause/Zeldes 2016).[2] Bereits in den ersten fünf Kapiteln nutzen wir Heische-Formeln mit *man* als Beispiel zur Veranschaulichung der jeweiligen Arbeitsphasen und resümieren im Anschluss an Kapitel 6 die wichtigsten Erkenntnisse.

2 Die Tools *INCEpTION* und *ANNIS* können kostenfrei unter https://inception-project.github.io/ (Stand: 3.12.2025) sowie https://corpus-tools.org/annis/download (Stand: 3.12.2025) heruntergeladen werden.

2. Datensammlung

Zu Beginn der Projektarbeit bestand der erste Schritt in der Erschließung der für den Korpusaufbau relevanten Textgrundlagen als durchsuchbare TXT-Dateien. Die Datenbasis beider Projekte ist in Tabelle 1 dargestellt.[3]

Projekt	**Zeitraum**	**Anzahl der Texte je Gattung**	**Anzahl der Token im kürzesten/ längsten (Sprech-)Text**	**Anzahl der Token insgesamt und im Durchschnitt**
Teilprojekt II *Das Pronomen ‚man' in der Diachronie des Deutschen*	15.–17. Jh.	Pesttraktate (12)	n = 1.867/5.928	n = 77.590 Ø = 6.466
		Bäderkunden (7)	n = 7.975/16.180	n = 76.604 Ø = 10.943
Teilprojekt IV *Personenreferierende Pronomen in Dramen*	17.–19. Jh.	Dramen (31)	n = 4.153/54.126	n = 754.783 Ø = 24.348

Tab. 1: Datenbasis der beiden Teilprojekte nach Zeitraum, Gattung und Tokenanzahl

Der Tabelle ist zu entnehmen, dass die beiden präsentierten Korpora mit je 19 und 31 Texten eine unterschiedlich große Datenmenge umfassen. Die Pesttraktate sind im Durchschnitt am kürzesten, die Dramen zeichnen sich dagegen durch die längsten Texte aus. Die Länge der einzelnen Texte variiert insgesamt stark: von 1.867 Token (im *man*-Projekt) bis 54.126 Token (im Dramen-Projekt). Die Korpuszusammensetzung wird im Folgenden in Verbindung mit der Beschreibung der Textsammlung genauer erläutert.

Für das Dramen-Projekt lagen die Dateien bereits aus unterschiedlichen Quellen vor: So konnte für die Dramen von A. Gryphius, der die Epoche des Barocks repräsentiert, auf das aus zehn Stücken bestehende Korpus zurückgegriffen werden, das 2017–2020 im Rahmen des DFG-Projekts *Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive* erstellt und im JSON- bzw. CLS-Format veröffentlicht wurde (vgl. Imo et al. 2020; im Folgenden Gryphius-Projekt).[4] Für die Texte der drei übrigen Autoren wurde das online frei zugängliche *German Drama Corpus* (*DraCor*, vgl. Fischer et al. 2019) herangezogen, welches über 700 deutsche Dramen vom 15.–20. Jahrhundert im TXT- bzw. TEI-Format sowie bestimmte Metainfor-

3 Eine detaillierte Übersicht der konkreten Texte findet sich in Duve (2024) und Schick (2024). Im *man*-Projekt werden neben den hier genannten Textgattungen weitere zum Vergleich hinzugezogen. Diese sind z. T. bereits als Korpora vorhanden (neue Zeitungen, Hexenverhörprotokolle) oder liegen in Form einzelner TXT-Dateien vor, die in Excel annotiert und nicht als eigenständiges Korpus aufbereitet werden (Komödien, Musterdialoge in Fremdsprachenlehrwerken). Als Token gelten in beiden Projekten alle deutschen sowie fremdsprachlichen Wörter und Satzzeichen und als ‚Sprechtext' ist die Figurenrede (ohne Regiekommentare) zu verstehen.

4 Auf das Gryphius-Korpus kann über das *TALAR – Tübingen Archive of Language Resources* zugegriffen werden: https://fdat.uni-tuebingen.de/records/hshq9-w0g79 (Stand: 3.12.2025).

mationen zu den entsprechenden Autoren und Werken enthält.[5] Daraus konnten stellvertretend für die Epochen der Aufklärung, des Sturm und Drang sowie der Klassik insgesamt 21 Dramen der Autoren G. E. Lessing (6 Dramen), F. Schiller (8 Dramen) und J. W. v. Goethe (7 Dramen) ausgewählt werden. Die Auswahl fiel auf die bekanntesten Werke der Autoren, die einerseits die jeweilige Epoche sprachhistorisch und -stilistisch am besten widerspiegeln und andererseits eine adäquate Annäherung an die Rekonstruktion des mündlichen Sprachgebrauchs dieser Zeit zulassen. Die quantitative Ausgewogenheit der Teilkorpora war dabei aufgrund der vorrangig qualitativen Untersuchungsziele des Projekts kein gewichtiger Faktor.

Das Korpus des *man*-Projekts ist in zeitliche Abschnitte von 50 Jahren unterteilt und darüber hinaus nach räumlichen Parametern strukturiert. Es wurde pro Zeitabschnitt jeweils ein Text aus der ostmittel- und ostoberdeutschen Region einbezogen, da beide Regionen die nhd. Schriftsprache maßgeblich prägten. Die Texte sollten außerdem möglichst ähnlich im Umfang sein, daher wurde für den quantitativen Ausgleich teilweise ein zweiter Text ergänzt. Alle Schriftstücke lagen zunächst nur als PDF-Bilddigitalisate vor, welche über die Datenbank *Deutsche Fach- und Wissenschaftssprache bis 1700* (vgl. Klein et al. 2017) erschlossen wurden.[6] Diese Datenbank bündelt Metadaten und Informationen über verfügbare Digitalisate von Texten, „die für die frühe Geschichte der deutschen Fach- und Wissenschaftssprachen von Bedeutung sind“ (ebd.). Darunter fallen auch Pesttraktate und Bäderkunden, die dem Sachbereich ‚Medizin‘ zugeordnet werden. Die Datenbank verweist auf die jeweiligen Bibliotheken, bei denen die Digitalisate der Drucke frei verfügbar vorliegen. Über diesen Weg wurden die zugehörigen PDF-Dateien heruntergeladen.

Daraufhin wurde im *man*-Projekt mit der Software *Transkribus* (vgl. Kahle et al. 2017) eine automatische Texterkennung (OCR) durchgeführt.[7] *Transkribus* ist eine KI-gestützte Plattform, die Tools zur Digitalisierung, Transkription und Bearbeitung von historischen Drucken oder Handschriften bereitstellt (vgl. ebd., S. 463). Der Zugang ist bis zu einem bestimmten Textkontingent frei, für die Verarbeitung größerer Textmengen und für bestimmte Texterkennungsmodelle und Tools muss eine Nutzungslizenz erworben werden. Ein Kernangebot von *Transkribus* ist die Bereitstellung von Texterkennungsmodellen, die für unterschiedliche Textarten (Sprachen, Zeiträume, Druckarten) trainiert wurden. Neben den von *Transkribus* erstellten Modellen, sind auch Modelle anderer NutzerInnen verwendbar. Auch das Trainieren eigener Modelle ist möglich, lohnte sich aber für dieses Projekt angesichts der kleinen Textmenge nicht. Ein Problem bei der Suche nach einem passenden Modell war, dass die meisten spezifischeren, an Konventionen und Druckpraktiken der frühen Neuzeit angepassten Modelle Ligaturen, diakritische Zeichen oder Abkürzungen wie Nasalstriche auflösen, die aber gerade für historische Korpora relevante Informationen darstellen und daher auch in diesem Projekt erhalten bleiben sollten. Daher wurde das Modell *Transkribus Print M1* genutzt, welches an Drucken in verschiedenen europäischen Sprachen erprobt wurde (vgl. Transkribus Print Mehrsprachig 2023). Bei diesem Modell wird auch das Layout des jeweiligen Digitalisats automatisch erkannt und kann ggf. angepasst werden, um Zeilen- und Seitenumbrüche originalgetreu wiederzugeben (vgl. Abb. 1).

5 Das *DraCor* ist über https://dracor.org/ger (Stand: 3.12.2025) zu erreichen.

6 Die Datenbank findet sich unter: https://fachtexte.kallimachos.de/ (Stand: 3.12.2025).

7 Die *Transkribus*-Plattform ist unter folgender URL zu erreichen: www.transkribus.org/ (Stand: 3.12.2025).

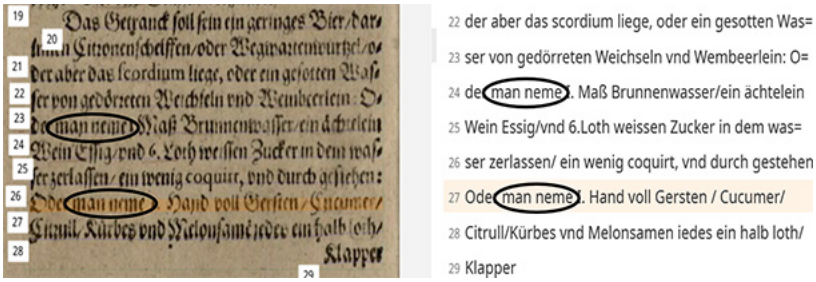

Abb. 1: Texterkennung eines Pesttraktats von C. H. Ayrer (1607) – Ansicht in *Transkribus* © Münchener Digitalisierungszentrum

Da das Modell eine Fehlerrate von 2,2% der Gesamtzeichen aufweist, mussten die entstandenen Transkripte noch einmal Korrektur gelesen werden. Die tatsächliche Fehlerrate variierte sehr stark in Bezug auf die Qualität und das Layout der Drucke in den Digitalisaten. Demzufolge ist bei der Planung eines solchen Vorhabens nur bedingt im Voraus bestimmbar, wie viel Zeit die Revision eines Datums in Anspruch nehmen wird. Von Vorteil ist jedoch, dass die Textkorrektur direkt in *Transkribus* durchgeführt werden kann. Sie wird dadurch erleichtert, dass z. B. Parallelstellen durch eine farbige Markierung gut erkennbar sind, sodass bei der Revision effizient vorgegangen werden kann (vgl. Abb. 1). Zudem ist es möglich, Dokumente per Link zu teilen und so gemeinsam Korrekturen vorzunehmen. Am Ende der Korrektur können die Texte als TXT-Dateien exportiert werden.

3. Bearbeitung der Daten

Die Daten beider Korpora wurden im Anschluss an die Vorverarbeitung in Anlehnung an den im Gryphius-Projekt konzipierten Workflow für die Annotation vorbereitet (vgl. Imo et al. 2020). Jeder Volltext wurde zunächst einzeln in ein im Rahmen des Gryphius-Projekts entwickeltes webbasiertes Tokenisierungstool eingefügt.[8] Das Tool segmentiert den Text automatisch in einzelne Token und gibt ihn im Text Corpus Format (TCF)[9] heraus, das die folgende Annotation in *INCEpTION* ermöglicht. Dafür erhalten alle Interpunktions- bzw. Sonderzeichen und Wörter eine eigene Token-ID (vgl. Abb. 2). Am Beispiel des Tokens mit der ID *"w394b">6.Loth<* ist jedoch zu erkennen, dass einige Zeichenkombinationen, z. B. die ohne zwischenstehende Leerzeichen, vom Programm fälschlicherweise als ein Token verarbeitet werden können, was je nach Forschungszweck einer manuellen Korrektur bedarf.[10] Bei diesem Zwischenschritt muss v. a. für ältere, weniger an einer Norm orientierte Texte genug Zeit eingeplant werden, da sich anderenfalls Probleme bei den Annotationen ergeben und z. B. im Falle der Lemmatisierung die Lemmata nicht tokengenau vergeben werden können.

8 Das frei verfügbare Tokenisierungstool ist unter folgender URL abrufbar: https://gryphius.sprache-interaktion.de/workflow/tokenizer_v4.html (Stand: 3.12.2025).

9 Eine Beschreibung des TCF erfolgt unter: https://weblicht.sfs.uni-tuebingen.de/englisch/tutorials/html/index.html (Stand: 3.12.2025).

10 Eine prozentuale Fehlerrate des Tools ist in der Dokumentation nicht aufgeführt und wurde in den Projekten nicht separat bestimmt.

Abb. 2: Heische-Formeln im Pesttraktat von C. H. Ayrer (1607) – Ansicht im Tokenizer

Außerdem werden auf Basis der Zeilenvorgaben des jeweiligen Originaltextes die entsprechenden Tokenkombinationen in Sätze eingeteilt. Das bedeutet, dass einer Zeile eine *sentence*-ID zugeordnet wird. Zu beachten ist dabei, dass dadurch sowohl ein einzelner Satz als auch mehrere Sätze unter eine *sentence*-ID gefasst werden können. Dies hat einerseits den Vorteil, dass bspw. in Dramentexten literarische Stilmittel wie Versprünge (Enjambements), die sich z. T. über mehrere Verse (Zeilen) erstrecken, erhalten bleiben und in der Annotationsansicht von *INCEpTION* gleichermaßen wiedergegeben werden. Andererseits müsste an dieser Stelle wie bei der Tokensegmentierung eine manuelle Nachkorrektur erfolgen, wenn die konkrete Satzanzahl relevant ist oder syntaktische Fragestellungen im Fokus stehen.

Das herausgegebene Tokenisierungsergebnis muss schließlich kopiert und in einem Texteditor (z. B. *Notepad++*, vgl. Ho 2003)[11] als Datei gespeichert werden, damit es im Webprogramm des Deutschen Textarchivs zur linguistischen Analyse historischer Texte *Cascaded Analysis Broker* (*DTA CAB*, vgl. Jurish 2012)[12] weiterbearbeitet werden kann. Dieses Programm ist wiederum für die orthografische Normkorrektur, die Lemmatisierung und das Part-of-Speech-Tagging (Wortartenzuweisung, im Folgenden POS-Tagging) der einzelnen Token, also die Ergänzung der eingespeisten Textvorlage um erste morphologische und morphosyntaktische Annotationsparameter, verantwortlich.[13] Die Verarbeitung kann dann je nach Textmenge einige Zeit (i. d. R. ein paar Minuten) in Anspruch nehmen. Nach aktuellem Stand bearbeitet der *DTA CAB* Dateien nur bis zu einer Größe von 1 MB. Bei größe-

11 Das Programm ist als Freeware unter https://notepad-plus-plus.org/ zu finden (Stand: 3.12.2025).

12 Auf den *DTA CAB* kann unter folgender URL zugegriffen werden: https://deutschestextarchiv.de/public/cab/ (Stand: 3.12.2025).

13 Sofern die Tokenisierung eines Textes im Vorfeld durchgeführt wurde, ist diese Option in der Maske des DTA-Webservice vor der Weiterverarbeitung zu entfernen, um Datenkontaminationen zu vermeiden.

ren Datenmengen kann allerdings der Programmdokumentation zufolge das DTA für weitere Informationen kontaktiert werden.[14] Abschließend muss das Ergebnis erneut als TCF-Datei gespeichert werden, welche die Grundlage der nächsten Arbeitsphase bildet.

4. Annotation

Nach der automatisierten Vorverarbeitung der Korpusdaten wurden diese in das Annotationsprogramm *INCEpTION* eingespeist. Für beide Projekte wurde in Absprache mit der IT der Universitäten jeweils eine eigene Webinstanz aufgesetzt, die das rechnerunabhängige Zugreifen auf die Daten und so die parallele Datenbearbeitung durch mehrere annotierende Personen über separat erstellte NutzerInnenaccounts zulässt. Es ist aber prinzipiell auch möglich, die Software lokal auf dem eigenen Rechner zu installieren.

In der jeweiligen Webinstanz wurde zunächst im Admin-Account ein Annotationsprojekt erstellt, in das die zuvor aufbereiteten TCF-Dateien hochgeladen wurden. Daraufhin wurden dort projektintern ausgearbeitete Layer (Annotationsebenen) angelegt, die anschließend mit den zugehörigen Tagsets (Bündel von Annotationskürzeln) verknüpft wurden.[15] Ein Layer kann dabei grundsätzlich auf unterschiedliche formale oder funktionale Codierungsziele ausgerichtet sein. Das entsprechende Tagset erlaubt – je nach Bedarf – als freies Tagset das eigenständige Befüllen des Annotationsfeldes (bspw. im Falle einer erforderlichen orthografischen Korrektur), während es als geschlossenes Tagset (bspw. auf der POS-Ebene) eine Vorauswahl an Merkmalsausprägungen der Annotationskategorie enthält. Für die formalen Layer sind in *INCEpTION* voreingestellte Tagsets vorhanden. So sind etwa für das POS-Tagging die Kodierungen des STTS (Stuttgart-Tübingen-Tagset, vgl. Schiller/Teufel/Stöckert 1995) hinterlegt, und es ist bereits eingestellt, dass die Annotation tokengenau erfolgt. Das vorbereitete Annotationsprojekt wurde anschließend für alle NutzerInnenaccounts zur Bearbeitung freigegeben.

Da die Texte beider Projekte eine automatische Vorannotation im *DTA CAB* durchlaufen haben, umfasste der erste Bearbeitungsschritt die parallele manuelle Korrektur jedes Textes auf den Ebenen i) orthografische Norm und ii) Lemma auf der Basis zuvor entwickelter projektspezifischer Guidelines.[16] Zeitgleich zur darauffolgenden POS-Korrekturphase wurde vor dem Hintergrund erster laufender Analysen mit der linguistischen Annotation ausgewählter Phänomene begonnen. So wurden im Dramen-Projekt für die ersten Texte Personalpronomen nach grammatischen Kategorien wie Kasus, Numerus, Genus etc. sowie nach semantisch-funktionalen Parametern wie Referenztyp oder Vorliegen einer Höflichkeitsform annotiert (vgl. das Token *man* in Abb. 3). Auf das Spektrum der Possessiv- und Demonstrativpronomen soll in einer späteren Projektphase ebenfalls genauer fokussiert werden. Im *man*-Projekt wurden Elemente der temporalen und lokalen Deixis annotiert,

14 Die Programmdokumentation des *DTA CAB* steht unter https://kaskade.dwds.de/~moocow/software/DTA-CAB/doc/html/DTA.CAB.WebServiceHowto.html (Stand: 3.12.2025) zur Verfügung. Im Dramen-Projekt wurden einige umfangreichere Werke im Texteditor Notepad++ nach einem eigenständig erarbeiteten Vorgehen händisch in mehrere Teildateien segmentiert und nach den beschriebenen Arbeitsschritten wieder zusammengesetzt. Dies kann hier allerdings aus Umfangs- und Komplexitätsgründen nicht im Detail aufgeschlüsselt werden.

15 Die genaue Vorgehensbeschreibung findet sich im User-Guide unter: https://inception-project.github.io/releases/33.5/docs/user-guide.html#_introduction (Stand: 3.12.2025).

16 Sowohl im Dramen- als auch im *man*-Projekt wurde sich zu Korrekturbeginn an den Guidelines zur formalen Annotation des o. g. Gryphius-Projekts (vgl. Imo et al. 2020) orientiert, da die Daten mit denselben Tools bearbeitet wurden. Es ist jedoch i. d. R. notwendig, vorhandene Guidelines mit Blick auf die eigenen Untersuchungsziele und die Datenspezifika aus der Praxis heraus anzupassen, was in beiden Teilprojekten innerhalb der jeweiligen Annotationsschleifen umgesetzt wurde.

da diese Hinweise auf Referenztypen liefern können. Zudem wurde in beiden Projekten eine Pilot-Codierung von Verben (Verbklasse, Bedeutungsgruppe u. Ä.) basierend auf Guidelines vorgenommen, die innerhalb der o. g. Forschungsgruppe *Praktiken der Personenreferenz* projektübergreifend konzipiert wurden (vgl. das Token anzůnde in Abb. 3). Im Zuge dessen können in beiden Datensets weiterhin nicht nur tokenspezifische Informationen annotiert werden, sondern auch Relationen zwischen einzelnen Token, die den Zusammenhang zwischen ihnen aufzeigen, wie die Verknüpfung zwischen *man* und *anzůnde*, die das Label <Heische Formel> trägt (vgl. Abb. 3). Die Annotation von Relationen ist eine Besonderheit in *INCEpTION*, die in anderen Annotationstools nicht immer vorhanden ist.

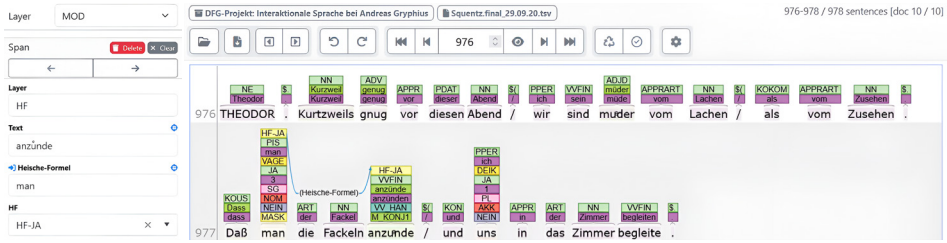

Abb. 3: Annotation einer Heische-Formel im Drama Peter Squentz von A. Gryphius (1657) – Ansicht in *INCEpTION*

Sowohl die Korrektur der formalen Merkmale als auch die Annotation linguistischer Kategorien wird jeweils von zwei AnnotatorInnen unabhängig voneinander im eigenen Account vorgenommen und daraufhin im Kurationsprozess miteinander verglichen.[17] Das Programm zeigt im Admin-Account innerhalb dieses Bearbeitungsschrittes durch farbige Markierungen der Annotationsfelder (Nicht-)Übereinstimmungen an. Diese Option wurde zum einen dafür genutzt, nach ersten Annotationen die Guidelines anzupassen und zu verfeinern, um die Fehlerrate zu verringern. Zum anderen werden die jeweiligen Annotationsversionen auf diesem Weg vom Admin regelmäßig korrigiert und wieder zu einer Version zusammengeführt, in der die nächste Annotationsetappe erfolgt. Die kuratierten Dateien sollten dann im TSV-Format exportiert werden, da nur in diesem Format alle eigenständig hinzugefügten Annotationen erhalten bleiben.

5. Auswertung

In einem letzten Schritt werden die annotierten Texte in das Korpustool *ANNIS* (vgl. Krause/Zeldes 2016) geladen, das die Suche nach sowie den Export und die Visualisierung von Belegen sprachlicher Phänomene ermöglicht. Dafür werden die entsprechenden Dateien aus *INCEpTION* im TSV-Format exportiert und mithilfe des Tools *Pepper* (vgl. Zipser/Romary 2010) in das von *ANNIS* benötigte GraphML-Format konvertiert.[18] *Pepper* ist ein *Java*-basiertes Programm, welches in der Command-Line eines Rechners bedient werden kann. Es ermöglicht die Umwandlung von linguistischen Korpora unterschiedlichster Ausgangsformate in andere Dateiformate. Es wurde jedoch v. a. mit dem Ziel entwickelt, die Konvertierung in das für *ANNIS* spezifische GraphML-Format zu ermöglichen. Während des Konvertierungsprozesses können auch einige Anpassungen der Daten wie Tagging-

17 Ein Inter-Annotator Agreement (vgl. Lemnitzer/Zinsmeister 2015, S. 61) wurde für die formalen Annotationsebenen nicht bestimmt. Für die funktionalen und damit fehleranfälligeren Kategorien wird dies angestrebt, allerdings wird die Annotation dieser Kategorien erst zum Ende der Projektlaufzeit abgeschlossen sein.

18 Das Tool *Pepper* ist unter https://corpus-tools.org/pepper/ (Stand: 3.12.2025) zu finden.

korrekturen bzw. -ergänzungen oder die Zusammenführung von Datensätzen vorgenommen werden.

ANNIS eignet sich als Infrastruktur zur Visualisierung und Auswertung von historischen Korpora, da Sonderzeichen wie Ligaturen oder diakritische Zeichen problemlos angezeigt werden. Für pragmatische Fragestellungen ist die Suchfunktion in *ANNIS* von Vorteil, da über die Suchsyntax (*ANNIS Query Language*)[19] neben einzelnen Token auch nach komplexen Phänomenen sowie auf unterschiedlichen linguistischen Ebenen gesucht werden kann. So können etwa V2- und VL-Heische-Formeln mit *man* ermittelt werden, indem nach dem Pronomen auf der Lemma-Ebene und nach finiten Verben (Voll-, Modal- und Auxiliarverb) im Konjunktiv I im Kontext von sechs Token mit der folgenden Abfrage gesucht wird:

LEMMA="man" ^1,6 POS=/(VVFIN|VMFIN|VAFIN) / _=_ MOD=/M_KONJ1/

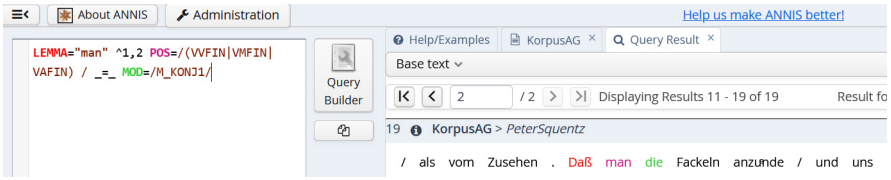

Abb. 4: Suche nach einer Heische-Formel in der Komödie *Peter Squentz* von A. Gryphius (1657) – Ansicht in *ANNIS*

Der relativ weit gesetzte Tokenabstand zwischen *man* und Verb dient der Ergebniserfassung der Fälle mit VL-Stellung, in denen komplexe NPs enthalten sind. Eine andere Möglichkeit stellt die Suche nach Relationen dar, wie z. B. nach der in *INCEpTION* veranschaulichten Relation < Heische-Formel >, die durch *ANNIS* angemessen dargestellt werden können. Die Belege können daraufhin im CSV-Format als KWICs oder mit den jeweils relevanten Annotationen exportiert werden. Dabei bietet das lokal installierbare Tool *Annimate* (vgl. Stemmler/Zemann i. d. B.) eine vereinfachte Infrastruktur für den Export.[20] Nach Projektabschluss werden die Korpora beider Teilprojekte voraussichtlich über eine öffentlich zugängliche *ANNIS*-Instanz für die Forschungsgemeinschaft verfügbar gemacht.[21]

6. Fallstudie: Heische-Formeln mit *man*

Im heutigen Sprachgebrauch finden sich sogenannte Heische-Formeln wie

(1) Man erinnere sich, daß…

oder

(2) Man nehme X (vgl. Zifonun/Hoffmann/Strecker 1997, S. 665; auch Duve/Schick eingereicht)

kaum noch wieder. Sie sind oft an spezifische (historische) Textgattungen und damit auch an spezifische Handlungsanforderungen gebunden (vgl. (1) für Zeitungen, (2) für Rezepte) und setzen sich im Gebrauch mit dem impersonalen Pronomen *man* in ihrer formalen

19 Genaue Informationen dazu können unter https://korpling.github.io/ANNIS/4/user-guide/aql/ (Stand: 3.12.2025) abgerufen werden.

20 *Annimate* kann unter https://github.com/matthias-stemmler/annimate (Stand: 3.12.2025) kostenfrei heruntergeladen werden.

21 Als Alternative dazu werden die Korpusdateien in unterschiedlichen Formaten zur Verfügung gestellt, sodass sie über eine lokal installierte Version von *ANNIS* oder mit anderen Tools analysiert werden können.

Grundstruktur aus *man + Verb in 3. Pers. Sg. Konj. I* zusammen.[22] Auf funktionaler Ebene werden Heische-Formeln Aufforderungsakten zugeordnet und stehen darin Imperativkonstruktionen nahe (vgl. Jäger 1970, S. 108), wobei die genaue Ausgestaltung der direktiven Kraft durch das weite Referenz- bzw. Funktionsspektrum des Pronomens und des Konjunktivs I sehr variabel und kontextsensitiv ist (vgl. Duve/Schick eingereicht). Vor diesem Hintergrund ist es interessant, zu überprüfen, ob die Struktur in früheren Sprachstufen frequenter war und/oder in einem breiteren Funktionsspektrum genutzt wurde. In diesem Zusammenhang hat eine erste empirische Untersuchung dieses Phänomens in Dramen des 16.–19. Jahrhundert gezeigt, dass die Konstruktion in dieser Gattung als Positionierungsressource abhängig von der sozialen Konstellation der SprecherInnen genutzt wird, um direktive Sprechakte zu formulieren, ohne u. a. soziale Hierarchien zu verletzen (vgl. Duve/Schick eingereicht). In der vorliegenden Fallstudie wird – wenn auch nur exemplarisch – der Blick auf die beiden weiteren Textgattungen des *man*-Projekts ausgeweitet.

Dafür wurde zunächst in beiden Korpora (vgl. Tab. 1) wie oben beschrieben nach Heische-Formeln mit *man* gesucht. Da für die Analyse des funktionalen Spektrums dieser Heische-Formeln die Belege in ihrem weiteren situativen Kontext aus qualitativer Sicht betrachtet werden müssen, wurde für den Export der Belegkollektion in *ANNIS* die größtmögliche Tokenzahl für den Kontext angegeben. Dass dieser in *ANNIS* besonders weit eingestellt werden kann, ist ein weiterer Vorteil des Tools für pragmatische Fragestellungen. Unentbehrlich blieb dabei allerdings die parallele Arbeit mit den Volltexten, da die inhaltliche Einbettung der Belege für ihre Interpretation zentral ist. Bei der interaktional ausgerichteten Analyse eines Dramenbelegs ist etwa das vorherige und nachfolgende Geschehen relevant, um die fokussierte Figurenkonstellation zu verstehen. Bei Rezeptbeschreibungen in den medizinischen Textsorten ist wiederum die Einordnung in Kapitel von Bedeutung, um zu rekonstruieren, ob die Heische-Formel sich an Kranke oder praktizierende Ärzte richtet.

Da die in Kapitel 3 dargestellte Annotation von Relationen in den Projekten noch aussteht, wurde in *ANNIS* nach dem Lemma *man* in Kombination mit einem Verb im Konjunktiv I im Abstand von sechs Token gesucht (vgl. Kap. 5). Nach Durchsicht und Eliminierung von Fehltreffern dienen 182 Belege (Pesttraktate n = 49, Bäderkunden n = 40, Dramen n = 93) als Datenbasis dieser Fallstudie. Die quantitative Übersicht in Abbildung 5 zeigt die relative Häufigkeit von *man* in Heische-Formeln pro 10.000 Token des Subkorpus der jeweiligen Gattung. Zwischen den Gattungen ist dabei ein deutlicher Unterschied erkennbar. Die Dramen weisen weniger Belege auf als die anderen beiden Gattungen[23] und es ist zu sehen, dass die Dichte an Heische-Formeln in den Pesttraktaten höher ist als in den Bäderkunden. Bei den Dramen zeigt sich zudem eine große Varianz zwischen den Autoren, wobei Heische-Formeln mit *man* in den Werken von Gryphius mit 3 Belegen/10.000 Token am häufigsten vorkommen (vgl. Abb. 6). Da sich die beiden Korpora nur in der zweiten Hälfte des 17. Jahrhundert überschneiden (vgl. Tab. 1), kann keine Aussage über den diachronen Verlauf getroffen werden. Auch innerhalb der Gattungen sind keine

22 Auch wenn diese Formeln mit unterschiedlichen Personal- und Indefinitpronomen realisiert werden können (vgl. Zifonun/Hoffmann/Strecker 1997, S. 663 f.), legen wir den Fokus vor dem Hintergrund unserer Projektarbeit auf den übergreifend relevanten Einsatz des impersonalen Pronomens *man* im Rahmen dieser spezifischen Struktur.

23 Die Anzahl der Belege gilt ausschließlich für den Sprechtext der Figuren, der für das interaktional ausgerichtete Analysevorgehen des Dramen-Projekts grundlegend ist. Sie würde jedoch mit hoher Wahrscheinlichkeit auch bei Einbezug von Paratexten wie Vor- bzw. Nachwort nicht signifikant höher sein, da innerhalb dieser Textpassagen direktive Handlungsformate nur in Ausnahmefällen wie der Adressierung der LeserInnen zu erwarten sind.

eindeutigen zeitlichen Entwicklungen im Einsatz der Heische-Formeln zu beobachten (vgl. beispielhaft Abb. 6 in Bezug auf die Dramen).

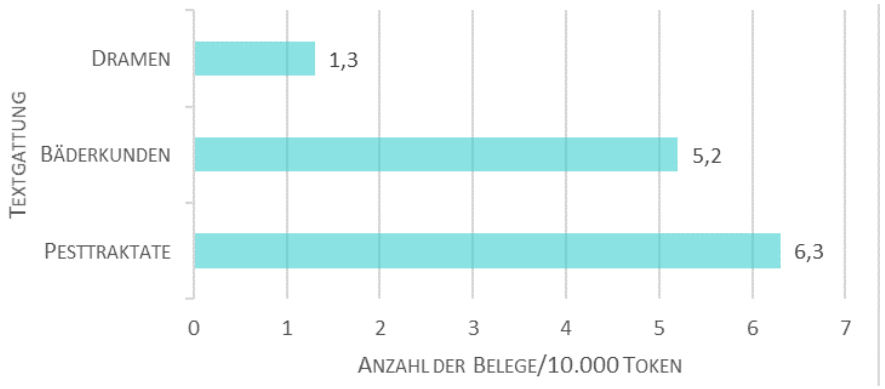

Abb. 5: Belege für *man* als Teil einer Heische-Formel pro 10.000 Token in den drei vertretenen Gattungen (n = 182)

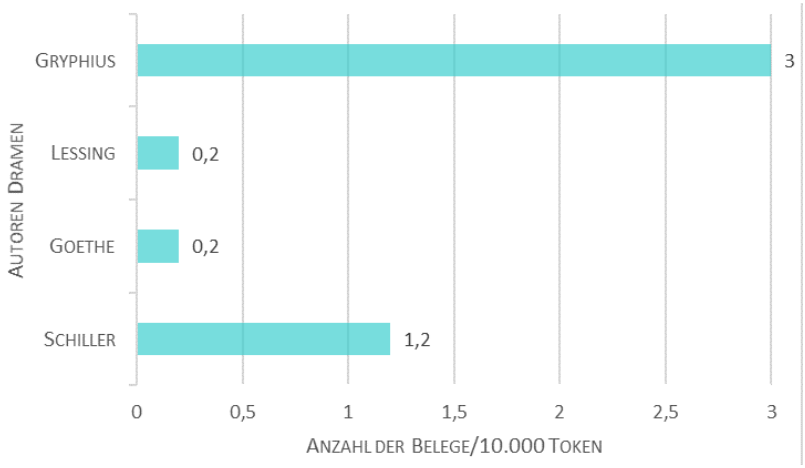

Abb. 6: Belege von *man* als Teil einer Heische-Formel pro 10.000 Token in den Dramen der jeweiligen Autoren (n = 93)

Im Folgenden wird exemplarisch anhand von zwei Textpassagen aufgezeigt, welche Funktionen Heische-Formeln mit *man* im Sprechtext der Dramen auf der einen und den medizinischen Gattungen auf der anderen Seite aufweisen. Bei Beispiel 3 handelt es sich um die Schlussszene der barocken Komödie *Peter Squentz* (1657) von A. Gryphius:

(3) **A. Gryphius – *Peter Squentz***
P. SQ. […] Gute Nacht Herr Kǒnig. Gute Nacht Frau Kǒnigin: gute Nacht Juncker / gute Nacht Jungfer / gute Nacht ihr Herren alle mit einander […].
THEODOR. Kurtzweils gnug vor diesen Abend / wir sind můder vom Lachen / als vom Zusehen. Daß man die Fackeln anzůnde / und uns in das Zimmer begleite.

In dieser Passage verabschiedet sich der Schreiber Peter Squentz, der für die Regieführung bei einem Theaterstück am königlichen Hofe verantwortlich war, nach der erfolgten Vorstellung von den ZuschauerInnen, zu denen u. a. der König Theodorus gehört. Daraufhin verkündet der König, dass die Adelsgesellschaft sich nach dem amüsanten Spektakel in ihre Gemächer zurückziehen möchte. In diesem Zuge äußert Theodorus die Heische-Formel *Daß man die Fackeln anzůnde / und uns in das Zimmer begleite*, welche im *man* + VL-Format innerhalb eines unabhängigen *dass*-Satzes gebildet wird (vgl. zu Heische-Formel-

Formaten in Dramentexten Duve/Schick eingereicht). Die darüber transportierte bindende Aufforderung gilt vor dem Hintergrund der sozialen Rolle, die der König repräsentiert, einer oder mehreren Gefolgsperson/en, deren Anwesenheit zwar nicht explizit erwähnt wird, jedoch kontextuell durch die Referenzleistung des impersonalen Pronomens *man* erschließbar ist. Durch die Verwendung der Heische-Formel mit offener Referenzleistung wird also deutlich, dass der König von der Erfüllung seines Willens ausgeht und dabei für ihn die konkreten Dienstleistenden keine Rolle spielen.

Wie die Studie von Duve/Schick (eingereicht) im Detail gezeigt hat, kommen Heische-Formeln in den innerhalb der beiden Teilprojekte untersuchten Dramentexten in je zwei V2- und VL-Realisierungsformaten vor, die zu unterschiedlichen Anteilen fünf direktive Handlungen, nämlich Irrelevanzkonditionale, Ratschläge, Ausrufe sowie (nicht-)bindende Aufforderungen, konstituieren (vgl. für erstere z. B. Schiller (1787): *Man nenn es Grille – Eitelkeit: gleichviel.*).[24] Mittels dieser Handlungen werden sowohl von Figuren eines höheren als auch niedrigeren sozialen Status verschiedene Willensbekundungen offengelegt. Dabei variieren die situativen Kontexte, in denen diese Willensbekundungen eingebettet sind, stärker als in den sich kontextuell ähnelnden Bäderkunden und Pesttraktaten, was am folgenden Beispiel im Ansatz verdeutlicht wird.

Der zweite ausgewählte Auszug entstammt einem Pesttraktat von C. H. Ayrer (1607) und ist Teil eines Kapitels, welches auf Getränke und Speisen eingeht, die für an der Pest Erkrankte empfohlen werden. Hier werden zwei Heische-Formeln im Rahmen einer Anleitung für die Zubereitung von gut bekömmlichen Getränken verwendet:

(4) **C. H. Ayrer – *Regiment und Ordnung vor den Seuchen der Pestilenz und Ruhr zu verwahren***
Oder man neme 1. Maß Brunnenwasser/ ein åchtelein Wein Essig/ vnd 6. Loth weissen Zucker in dem wasser zerlassen/ ein wenig coquirt, vnd durch gestehen: Oder man neme 1. Hand voll Gersten/ Cucumer/ Citrull/ Kůrbes vnd Melonsamē iedes ein halb loth/ Klapper oder klitzen Rosen/ Erbsichbeerlein jedes 1. Hand voll/ vnd lasse es in 2. Maß brunnē wasser sielen/ Tz letzt kann man ein wenig Zimmer daran thun/

Auch hier fungieren die V2-Heische-Formeln *man neme* und (*man*) *lasse sielen* als Aufforderungen, die allerdings sowohl an Kranke als auch an Wundärzte, die Kranke zu dieser Zeit versorgten (vgl. Eckkrammer 2015, S. 30), gerichtet sind. Das Pronomen *man* ermöglicht durch sein offenes Referenzpotenzial, dass die LeserInnen sich nach Bedarf von der Aufforderung angesprochen fühlen können. Bei der Erläuterung einer Zubereitungsoption im letzten Satz wird zusätzlich die Formulierung *kann man* verwendet, wobei das Modalverb *können* als Optionalitätsindikator aufzeigt, dass die vorausgehenden Heische-Formeln im Kontrast dazu eindeutig relevante Anweisungen kontextualisieren. Bemerkenswert ist außerdem, dass die sich hier wiederholende Konstruktion *man neme* als musterhafte Formel textübergreifend v. a. in den Pesttraktaten von Ayrer (1607), Strobelberger (1630) und Schorer (1667) vermehrt auftaucht (je 58,3%, 16,6% und 19% der Belege von Heische-Formeln mit *man* des jeweiligen Textes).

Insgesamt kommen alle Heische-Formeln (n = 89) sowohl in den Bäderkunden als auch in den Pesttraktaten in einem ähnlichen Kontext von Anleitungen bzw. Anweisungen vor. Diese können zum einen dazu dienen, konkrete Handlungsschritte zur Zubereitung von Medizin oder Bädern zu beschreiben oder Handlungsoptionen – etwa im Rahmen der Nut-

24 Für diese Studie wurden insgesamt 108 Belege aus 51 Dramentexten des 16.–19. Jahrhunderts ausgewertet. Die beiden anderen Textgattungen, die im *man*-Projekt analysiert werden, wurden dabei nicht einbezogen.

zung von Heilbädern – zu präferieren (vgl. Ryff 1549: *das man sich nit ubersettige mit speyß dermassen/ das der Magen darvon beschweret werde oder ein unwillen*). Die leicht höhere Belegzahl in den Pesttraktaten könnte sich daraus ergeben, dass dort anleitende Textteile einen größeren Stellenwert einnehmen, während in den Bäderkunden auch längere beschreibende oder theoretische Passagen vorkommen (vgl. Dammel/Duve eingereicht).

Das Funktionsspektrum in den beiden medizinischen Textgattungen ist demnach auf allgemeine Handlungsanweisungen beschränkt und ist damit eingegrenzter als in den untersuchten Dramentexten. Dies lässt sich v. a. auf die variierenden Drameninhalte sowie Beteiligungskonstellationen der Figuren zurückführen, die die Realisierung direktiver Sprechakte mit unterschiedlicher Verbindlichkeit relevant setzen (vgl. Duve/Schick eingereicht). Die höhere Frequenz von Heische-Formeln mit *man* in den anleitenden Texten legt die Tendenz nahe, dass sich für diese Gattungen schon früh typische und heute noch existente (wenn auch weniger gebräuchliche) Muster wie *man neme* (vgl. Ayrer 1607) herausgebildet haben. Zur Überprüfung dieser ersten Hypothesen wäre eine Ergänzung unserer Daten um die in den abgedeckten Zeiträumen fehlenden Gattungen (soweit vorliegend) notwendig. Dies wurde, wie oben erwähnt, anhand der im Rahmen des *man*-Projekts genutzten Dramentexte für das 16. Jahrhundert und die erste Hälfte des 17. Jahrhunderts zum Zwecke der Studie von Duve/Schick (eingereicht) bereits umgesetzt. In diesen Zeitabschnitten beträgt die relative Häufigkeit der Heische-Formeln nur 0,7 Belege pro 10.000 Token, was der Frequenz in den hier betrachteten Dramentexten nahekommt (vgl. Abb. 5, 6).

7. Fazit

Die Darlegung des Workflows zur programmgestützten Erstellung, Aufbereitung und Auswertung historischer Korpora sollte aufzeigen, wie diachrone, projektübergreifende Forschungsvorhaben ungeachtet unterschiedlicher Datengrundlagen ertragreich realisiert werden können. Dabei lag das Augenmerk darauf, inwieweit die von uns verwendeten Tools für die Arbeit (1) im Team, (2) mit pragmatischen Fragestellungen und (3) an historischen Daten besonders zielführend sind.

(1) Für eine kollektive Bearbeitung der Korpusdaten ist das von uns gewählte Programm *INCEpTION* insofern sehr gut geeignet, als es in Form einer Webinstanz den Zugriff auf die zu annotierenden Texte von verschiedenen Rechnern aus ermöglicht und es dem Administrator außerdem erlaubt, die zur Codierung benötigten Tagsets in einem Projekt einmalig anzulegen und sie allen Annotierenden gleichzeitig zugänglich zu machen. Des Weiteren erleichtert das Tool die Bündelung sowie Überprüfung mehrerer Annotationsergebnisse durch speziell dafür konzipierte Funktionen. (2) Ferner ist in *INCEpTION* bei der Auseinandersetzung mit pragmatischen Fragestellungen, wie den hier diskutierten Funktionen von Heische-Formeln mit *man*, die Möglichkeit der Erstellung von spezifischen Tagsets von Bedeutung, die konkret auf das eigene Forschungsinteresse zugeschnitten werden und mehrere linguistischen Ebenen umfassen können. In *ANNIS* sind wiederum die Auswertungs- und Darstellungsoptionen einfacher, aber auch komplexer Annotationen hilfreich, wie die für die Heische-Formeln codierte Beziehung zwischen pronominalen und verbalen Annotationsparametern. (3) Als Untersuchungsgrundlage stellen historische Daten aus den in Kapitel 2 und 3 beschriebenen Gründen in der Korpusaufbereitung meist eine Herausforderung dar. Es ist daher von Vorteil, zu ihrer Erschließung bestenfalls auf speziell für diesen Datentyp entwickelte und/oder daran erprobte Programme zurückzugreifen. Für die Vorhaben der hier diskutierten Projekte waren z. B. die Tools *Transkribus* und der *DTA CAB* unentbehrlich, um die jeweiligen Datensets für die Annotation vorzubereiten. Darüber hinaus war hinsichtlich der Textsegmentierung das für das Gryphius-Projekt imple-

mentierte Tokenisierungstool von Nutzen, da anderweitige Tokenizer ihre Analyse primär an orthografischen Zeichen ausrichten, die bspw. in frühneuzeitlichen Texten nur selten verwendet wurden und die demnach ungeeignete Anhaltspunkte für unsere Segmentierungsbedarfe geboten haben.

Anhand der Fallstudie zu Heische-Formeln mit *man* haben wir zudem illustriert, wie die projektübergreifend vorgenommenen Annotationen für konkrete Analysen fruchtbar gemacht werden können. Auffällig war dabei, dass Heische-Formeln v. a. in Pesttraktaten festzustellen waren, im Sprechtext der Dramen dagegen traten sie deutlich seltener auf. Aus qualitativer Sicht konnte ausgehend von den beiden exemplarisch diskutierten Auszügen aus einem Drama und einem Pesttraktat der Frühen Neuzeit das Funktionsspektrum der Konstruktion umrissen werden. Mit der Passage aus einer barocken Komödie wurde die Realisierung einer verbindlichen Aufforderung durch eine Heische-Formel mit *man* im Rahmen einer hierarchisch organisierten Interaktionssituation in den Blick genommen. Der Auszug aus einem Pesttraktat demonstrierte eine an die Leserschaft gerichtete Zubereitungsanleitung bzw. -anweisung, die über die darin eingesetzten Heische-Formeln kontextualisiert wurde und deren Ausführungsverbindlichkeit in erster Linie den (in-)direkt von der Krankheit betroffenen Personen gilt. Aufgrund des ähnlichen Gebrauchskontexts von Heische-Formeln innerhalb der Pesttraktate und Bäderkunden, mit denen bestimmte Adressatengruppen (Kranke, behandelnde Ärzte) allgemein im Handeln angeleitet werden, ist das Funktionsspektrum des Musters dort weniger ausdifferenziert als in Dramentexten mit variierenden Figurenkonstellationen und Interaktionssituationen im jeweiligen Geschehen. Dennoch lassen sich beide präsentierten Subfunktionen dem Dachbegriff direktiver Handlungen zuordnen, sodass sich in den untersuchten Datensets neben den gattungsbedingten Unterschieden durchaus auch Gemeinsamkeiten feststellen lassen.

Für weiterführende Forschungsvorhaben und die Ergänzung unserer Ergebnisse wäre z. B. ein Vergleich mit anderen Gebrauchstexten wünschenswert, denn Glaser (2002) weist im Hinblick auf die Textgattung ‚Kochrezepte' auf einen Anstieg im Einsatz von Heische-Formeln im 17. Jahrhundert hin, was u. a. in die Interpretation unserer quantitativen Beobachtungen neue Perspektiven einbringen würde.

Literatur

Dammel, Antje/Duve, Laura (eingereicht): Impersonale Pronomen als interpersonale Pronomen: Zur Perspektivierungsleistung von *man* im Zusammenspiel mit anderen Referenzformen in frühneuzeitlichen Bäderführern. In: Zeitschrift für germanistische Linguistik 26, 1 (Personenreferenz und Perspektive).

Duve, Laura (2024): Korpus. Teilprojekt 2: Referenzielle Praxis im Wandel: Das Pronomen *man* in der Diachronie des Deutschen. Forschungsgruppe Praktiken der Personenreferenz. https://tp2.forschungsgruppe-pronomen.de/korpus/ (Stand: 3.12.2025).

Duve, Laura/Schick, Valeria (eingereicht): „Man schaue und wundere sich." – Heische Formeln mit *man* als Positionierungsressourcen in Dramen des 16. bis 19. Jahrhunderts. In: Dammel, Antje/Imo, Wolfgang/Lanwer, Jens P. (Hg.): Pronomengebrauch und *stance taking*. Tübingen: Narr Francke Attempto.

Eckkrammer, Eva M. (2015): Medizinische Textsorten vom Mittelalter bis zum Internet. In: Busch, Albert/Spranz-Fogasy, Thomas (Hg.): Handbuch Sprache in der Medizin. (= Handbücher Sprachwissen 11). Berlin/München/Boston: De Gruyter, S. 26–46.

Fischer, Frank/Börner, Ingo/Göbel, Mathias/Hechtl, Angelika/Kittel, Christopher/Milling, Carsten/Trilcke, Peer (2019): Programmable corpora: Introducing DraCor, an infrastructure for the research

on European drama. In: Proceedings of Digital Humanities 2019: „Complexities“ (DH 2019). Utrecht, Utrecht University, S. 1–6.

Fürbeth, Frank (2012): Bäderdiskurse in den deutschsprachigen balneologischen Bestsellern des 16. Jahrhunderts (Paracelsus, Etschenreutter, Tabernaemontanus). In: Boisseuil, Didier/Wulfram, Hartmut (Hg.): Die Renaissance der Heilquellen in Italien und Europa von 1200 bis 1600. Geschichte, Kultur und Vorstellungswelt. Frankfurt a. M. u. a.: Lang, S. 193–212.

Glaser, Elvira (2002): Textmuster deutschsprachiger Kochrezepte im 19. und 20. Jh. In: Germanistik und Romanistik: Wissenschaft zwischen Ost und West. Materialien der internationalen wissenschaftlichen Konferenz „West-Ost: Bildung und Wissenschaft an der Schwelle des 21. Jahrhunderts“. Chabarovsk, 2002, S. 109–124.

Ho, Don (2003): Notepad++. https://notepad-plus-plus.org/ (Stand: 3.12.2025).

Imo, Wolfgang/Wesche Jörg/Müller, Melissa/Eggert, Lisa (2020): DGF-Projekt Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive. https://gryphius.sprache-interaktion.de/, https://gryphiusprojekt.wordpress.com/, https://gryphius.sprache-interaktion.de/workflow/ (Stand: 3.12.2025).

Jäger, Siegfried (1970): Indirekte Rede und Heischesatz. Gleiche formale Strukturen in ungleichen semantischen Bereichen. In: Moser, Hugo (Hg.): Studien zur Syntax des heutigen Deutsch. Paul Grebe zum 60. Geburtstag. (= Sprache der Gegenwart 6). Düsseldorf: Schwann, S. 103–117.

Jurish, Bryan (2012): Finite-state canonicalization techniques for historical German. Endliche Techniken zur Kanonikalisierung historischen deutschen Textes. Diss. Potsdam: Universität Potsdam.

Kahle, Philip/Colutto, Sebastian/Hackl, Günther/Mühlberger, Günter (2017): Transkribus – A service platform for transcription, recognition and retrieval of historical documents. 14th IAPR International conference on document analysis and recognition (ICDAR). Kyoto, Japan: IEEE: S. 19–24.

Klein, Wolf P./Breyl, Michael/Ahlers, Solveig/Kromer, Isabel/Schmid, Saskia (2017): Projektbeschreibung. https://fachtexte.kallimachos.de/Projektbeschreibung (Stand: 3.12.2025).

Klie, Jan-Christoph/Bugert, Michael/Boullosa, Beto/Eckart de Castilho, Richard/Gurevych, Iryna (2018): The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Zhao, Dongyan (Hg.): Proceedings of system demonstrations of the 27th international conference on computational linguistics (COLING 2018), Santa Fe, New Mexico. Association for Computational Linguistics, S. 5–9.

Krause, Thomas/Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 31, 1, S. 118–139.

Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik. Eine Einführung. 3., überarb. und erw. Aufl. Tübingen: Narr.

Schick, Valeria (2024): Unser Korpus. Teilprojekt 4: Personenreferierende Pronomen in Dramen: interaktionale und dramaturgische Funktionen sowie historischer Wandel von Barock über Aufklärung zu Sturm und Drang und Klassik. Forschungsgruppe Praktiken der Personenreferenz. https://tp4.forschungsgruppe-pronomen.de/unser-korpus/ (Stand: 3.12.25).

Schiller, Anne/Teufel, Simone/Stöckert, Simone (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart: Institut für maschinelle Sprachverarbeitung. www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1995.pdf (Stand: 3.12.2025).

SfS Tübingen (o. D.): TCF format. https://weblicht.sfs.uni-tuebingen.de/englisch/tutorials/html/index.html (Stand: 3.12.2025).

Transkribus Print Mehrsprachig (2023): https://readcoop.eu/de/modelle/transkribus-print-multilanguage-dutch-german-english-finnish-french-swedish-etc/ (Stand: 3.12.2025).

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7). Berlin/New York: De Gruyter.

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), May 2010, La Valette, Malta.

Kontaktinformation

Laura Duve
Universität Münster
Fachbereich 9 Philologie, Germanistisches Institut
Robert-Koche-Straße 29
48143 Münster
E-Mail: laura.duve@uni-muenster.de

Valeria Schick
Universität Hamburg
Fakultät für Geisteswissenschaften, Fachbereich SLM I, Institut für Germanistik
Von-Melle-Park 6
20146 Hamburg
E-Mail: valeria.schick@uni-hamburg.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.