ESV ERICH SCHMIDT VERLAG

PerspektivWechsel *oder:* Die Wiederentdeckung der Philologie

Band 1: Sprachdaten und Grundlagenforschung in der Historischen Linguistik

Herausgegeben von

Sarah Kwekkeboom
und
Sandra Waldenberger

ERICH SCHMIDT VERLAG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationablbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im
Internet über http://dnb.d-nb.de abrufbar.

Weitere Informationen zu diesem Titel finden Sie im Internet unter
ESV.info/978 3 503 16577 3

Gedrucktes Werk: ISBN 978 3 503 16577 3
eBook: ISBN 978 3 503 16578 0

Dieses Papier erfüllt die Frankfurter Forderungen
der Deutschen Nationalbibliothek und der Gesellschaft für das Buch
bezüglich der Alterungsbeständigkeit und entspricht
sowohl den strengen Bestimmungen der US Norm
Ansi/Niso Z 39.48-1992 als auch der ISO-Norm 9706.

Druck und buchbinderische Verarbeitung: Hubert & Co., Göttingen

Stefanie Dipper (Universität Bochum)

Variationen über ein Thema: statistisch bestimmte linguistische Ähnlichkeiten und Differenzen der Textzeugen des Anselm-Textes

In diesem Artikel geht es um das „Anselm-Korpus“[1], dessen Textgrundlage in Form diplomatischer Transkriptionen am Lehrstuhl von Prof. Dr. Klaus-Peter Wegera erstellt wurde.[2] Ein Großteil der Textzeugen lag bereits auf Microfiche oder als Fotografien aus dem Nachlass von Detmar Grubert vor und stellte ehemals die Grundlage eines letztlich nicht realisierten Dissertationsprojekts von Grubert dar. Im Rahmen des Projekts „*St. Anselmi Fragen an Maria* – (digitale) Erschließung, Auswertung und Edition der gesamten deutschsprachigen Überlieferung (14.–16. Jh.)“, das seit 2010 von der DFG gefördert wird, wurde das Korpus vervollständigt sowie sämtliche deutschsprachigen Textzeugen diplomatisch transkribiert, zunächst unter der Leitung von Klaus-Peter Wegera und Simone Schultz-Balluff (1. Projektphase), dann unter alleiniger Leitung von Simone Schultz-Balluff (2. Projektphase). In einem zweiten DFG-Projekt gleichen Namens unter meiner Leitung wurden computerlinguistische Verfahren zur weiteren Verarbeitung und Auswertung der Transkriptionen entwickelt. In diesem Artikel möchte ich Ergebnisse beider Projekte zusammenbringen.[3]

Die Textzeugen im Anselm-Korpus unterscheiden sich z.T. erheblich voneinander, haben jedoch alle dasselbe Thema: Anselm von Canterbury befragt Maria zu

---

1 S. Klaus-Peter Wegera: Interrogatio Sancti Anselmi de Passione Domini, deutsch. Überlieferung – Edition – Perspektiven der Auswertung. Nordrhein-Westfälische Akademie der Wissenschaften und der Künste – Geisteswissenschaften. Vorträge, G 445, Paderborn 2014.

2 Dieser Artikel hätte ohne die wertvollen (Vor-)Arbeiten meiner KooperationspartnerInnen aus der Germanistik – insbesondere Klaus-Peter Wegera, Simone Schultz-Balluff und Nina Bartsch – sowie meiner MitarbeiterInnen Marcel Bollmann, Julia Krasselt und Florian Petran nicht entstehen können. Ihnen danke ich vielmals, auch für Unterstützung speziell bei den Arbeiten zu diesem Artikel. Ebenfalls danken möchte ich den Herausgeberinnen Sarah Kwekkeboom und Sandra Waldenberger für ihre Kommentare und Verbesserungsvorschläge. Die Arbeit an diesem Artikel wurde unterstützt durch die DFG (Geschäftszeichen DI-1558/4).

3 Homepage des germanistischen Projekts:
http://www.ruhr-uni-bochum.de/wegera/sanktanselmus.html und
http://www.ruhr-uni-bochum.de/schultz-balluff/sanktanselmus.html;
Homepage des computerlinguistischen Projekts:
http://www.linguistics.ruhr-uni-bochum.de/anselm;
Alle URLs in diesem Artikel wurden am 13.08.2015 abgerufen.

den Leiden Christi. (1) und (2) zeigen zwei Versionen der ersten Frage Anselms und des Beginns der Antwort Marias aus zwei Handschriften, einer schwäbischen (1) bzw. thüringischen (2).[4]

(1) *Sant aunſhalm waz von herczen fro ſin frag hůb er an vnd ſprach zů ir ſag mir liebů frow wie was der anfank der marter dines lieben kindes vnſer frow ſprach do min kint het geſſen mit ſinen iungern vor ſiner marter daz Iungſt maſſ daz er nam vnd do ſi von dem tiſch vff ſtůnden do gieng iudaz ſcariot zů den Iuden der fůrſten vnd kam ainſ gedinges mit in vber ain daz er min kint wolt verrätten* (Stu, 96va,21–31)

„Sankt Anselm war von ganzem Herzen froh, er begann seine Befragung und sprach zu ihr: ‚Sage mir, liebe Frau, wie war der Anfang des Martyriums deines lieben Kindes?‘ Unsere Frau sprach: ‚Als mein Kind mit seinen Jüngern vor seinem Martyrium das letzte Mahl, das er nahm, gegessen hatte, und als sie von dem Tisch aufstanden, da ging Judas Ischariot zu den Fürsten der Juden und kam in einer Abmachung mit ihnen überein, dass er mein Kind verraten wollte.“

(2) *do ſprach ſente anſhelmus lybe vrouwe ſage myr dynes lyben kyndes marter ¶ Do ſprach vnſe vrouwe ¶ Do myn lybes kynt das obent brot hatte geſſen · myt ſynen iungeren vf geſtanden was noch dem eſſen ¶ Do lyf iudas czu den iuden vnd vorrit myn lybes kynt·*(B, 2v,4–14)

„Da sprach Sankt Anselmus: ‚Liebe Frau, berichte mir von deines lieben Kindes Martyrium.‘ Da sprach unsere Frau: ‚Als mein liebes Kind das letzte Abendmahl gegessen hatte, war er mit seinen Jüngern nach dem Essen aufgestanden. Darauf lief Judas zu den Juden und verriet mein liebes Kind.“

Die einzelnen Textzeugen lassen sich anhand von drei Hauptparametern in Klassen aufteilen: (i) anhand der sprachräumlichen Einordnung; (ii) anhand der zeitlichen Einordnung; (iii) anhand der Gattung Prosa vs. Vers. Am Lehrstuhl von Klaus-Peter Wegera wurden sämtliche deutschsprachigen Textzeugen manuell den entsprechenden Klassen zugeordnet.[5]

In diesem Artikel soll gezeigt werden, inwieweit sich diese manuelle Zuordnung der Textzeugen mit (einfachen) rein automatischen Verfahren replizieren lässt. Die sprachräumliche Klassifizierung korreliert z.B. mit einem bestimmten Lautstand, der sich zumindest in Teilen in der Verwendung bestimmter Buchstaben und Buchstabenkombinationen widerspiegelt. Daher soll hier versucht werden, die Textzeugen auf Basis ihrer Buchstaben und Buchstabensequenzen vollautomatisch in Klassen einzuteilen und diese anschließend mit der manuellen Zuordnung zu vergleichen.

---

4 *Stu* ist die Sigle der Handschrift, 96va,21–31 die genaue Stellenangabe des Belegs. Die Siglen der Überlieferungsträger sind in den Tab. 1 und 2 aufgelöst.

5 Das Ergebnis der Zuordnung ist in Form einer interaktiven Landkarte abrufbar unter http://www.ruhr-uni-bochum.de/wegera/texte_1.htm. Eine vervollständigte tabellarische Version davon steht unter http://www.ruhr-uni-bochum.de/wegera/Ueberlieferung_dt_1.htm. Ein Auszug aus dieser Tabelle sowie eine angereicherte Version der Landkarte finden sich auch in diesem Artikel (s. Tab. 1 und 2 sowie Abb. 5).

Zusätzlich soll auch eine Klassifizierung auf Basis von automatisch zugewiesenen Wortarten vorgenommen werden.

Im Folgenden wird zunächst das Korpus beschrieben (Abschnitt 1), dann die verschiedenen automatischen Verfahren (Abschnitt 2). Anschließend werden die Ergebnisse der manuellen mit den Ergebnissen der automatischen Verfahren verglichen (Abschnitt 3).

1. Das Anselm-Korpus

Die hier verwendete Version des Anselm-Korpus besteht aus 59 Textzeugen mit insgesamt rund 380.000 Wortformen und enthält damit alle zum aktuellen Zeitpunkt fertig transkribierten und kollationierten deutschsprachigen Textzeugen. Die Mehrzahl (49) sind Handschriften, der Rest (10) Drucke.

Tabellen 1 und 2[6] listen sämtliche (bekannten) deutschsprachigen Handschriften bzw. Drucke mit ihren Siglen und den am Lehrstuhl Wegera manuell vorgenommenen zeit- und sprachräumlichen Zuordnungen. Die zeitliche Zuordnung „15“ bedeutet „15. Jahrhundert“, „15.1“ bedeutet „1. Hälfte des 15. Jahrhunderts“. Die Markierungen mit „*“ und „**“ bedeuten „Überlieferungsträger verschollen“ bzw. „derzeit nicht zugänglich“, „***“ markiert noch nicht fertig kollationierte Transkripte. Von den so markierten Transkripten ist nur *sl* Teil des hier untersuchten Korpus und basiert auf einer Edition. Außerdem enthält *s1495* zwei Textzeugen des Anselm-Textes und ist im Korpus daher auf zwei Dokumente *s1495-T1* und *s1495-T2* verteilt.

Im Durchschnitt enthalten die Textzeugen rund 6.400 Wortformen. Allerdings schwanken die Umfänge sehr, die Standardabweichung beträgt fast 3.000. Die beiden umfangreichsten Textzeugen (*Ka* und *M*) enthalten beispielsweise 14.842 bzw. 10.273 Wortformen, die beiden kürzesten (Fragmente, erkennbar an den Siglen in Kleinschreibung: *hb* und *au*) 55 bzw. 145 Wortformen. Diese Unterschiede erschweren natürlich den Vergleich, und vor allem bei den extrem kurzen Fragmenten sind keine sinnvollen statistisch basierten Aussagen möglich. Trotzdem habe ich die kürzeren Textzeugen nicht aus den Untersuchungen ausgeschlossen, da solche Datengrößen bei historischen Sprachdaten nicht selten sind und es daher von Interesse ist, ihr statistisches Verhalten – unter den entsprechenden Vorbehalten – zu untersuchen. Ähnliches gilt für Textzeugen in Versform. Auch hier sind Unterschiede (z.B. in der Wortstellung) zu erwarten, die sich zu einem guten Teil auf die besondere Form zurückführen lassen. Auch diese Textzeugen schließe ich aus den Vergleichen nicht aus.

---

[6] Tab. 1 und 2 sind leicht modifizierte Versionen der Tabellen von http://www.ruhr-uni-bochum.de/wegera/Ueberlieferung_dt_1.htm (13.08.2014), ergänzt um Informationen zur Anzahl der Wortformen.

HANDSCHRIFTEN

Sigle	Dat.	Sprachraum	Prosa/ Vers	Aufbewahrungsort Signatur	Größe
au	15,1	obd., wobd., schwäb.	P	Augsburg, 2°Cod438	145
B	15	md., omd., thür.	P	Berlin, mgo183	5395
B2	15,2	md., wmd., rhfrk.	P	Berlin, mgq2025	9181
b	15,1	md., omd., obs.	V	Berlin, Fragm.4	2076
b2	15	ndd., ofäl., elbofäl.	P	Berlin, mgf736	746
B3	15,2	obd., oobd., mbair.	P	Berlin	6219
Ba	15,2	obd., nobd., nbair./ofrk.	P	Bamberg, Msc.Lit.176	6437
Ba2	15,2	obd., oobd., nbair./ofrk.	P	Bamberg, Msc.Lit.176	5958
Be	15,2	obd., wobd., hchalem.	P	Bern, Mss.h.h.X.50	8212
Br*	15,1		P		
D	14,2	ndd., mrk., smrk.	V	Dessau, GeorgHs.73.8	7477
D2	14,2	ndd., mrk., smrk.	V	Dessau, Hs.Georg.73.8°	7370
D3	15,2	md., omd., thür.	V	Dessau, Hs.Georg.24,8°(4°)	5709
D4	15,2	ob./md., wobd./wmd.,	P	Dessau, Hs.Georg.65.8°	6354
f	15,2	ndd., nndd., östl. nndd.	V	Fürstenwalde, oSignatur	2050
fb*	14	obd.	P		
H	16,1	md., omd., obs.	P	Halle, Qu.Cod.141	8445
hb	15,2	ndd., nndd., östl.nndd.	V	Hamburg, Cod.inscrin.17,Frgm.15	55
Hk	16	obd., oobd., mbair.	P	Heiligenkreuz, Cod.339	8718
Hk2**	15,2	obd., wobd., schwäb.	P		
Hz**	15,2	obd., oobd., bair, wohl mbair.,	P		
Ka	14,1	obd., wobd., hchalem./ndalem.	P	Karlsruhe, Cod.Donaueschingen116	14842
Kh	15	ndd., nndd.	V	Kopenhagen, Cod.Thott.109,4°	7137
Kn	14,1	obd., oobd., mbair.	V		
M	14	obd., oobd., mbair.	P	München, clm23371	10273
M2	15,1	obd., oobd., mbair.	P	München, cgm839	8740
M3	15,2	obd., oobd., mbair.	P	München, cgm485	7970
M4	15,2	obd., nobd., nbair./ofrk.	P	München, cgm484	8592
M5	15,2	obd., oobd., mbair.	P	München, cgm4698	4711
M6	15,2	obd., oobd., mbair.	P	München, cgm486	4639
M7	15,2	obd., nobd., mbair.	P	München, cgm473	4647
M8	15,2	obd., nobd., nbair.	P	München, cgm134	8462
M9	15,2	obd., oobd., mbair.	P	München, cgm4701	4748
M10	15	obd., oobd., mbair.	P	München, clm14945	4388
Me	15	obd., oobd., mbair.	P	Melk, Cod.55	4778
n	15,2	obd., nobd., nbair./ofrk.	P	Nürnberg, Cent.VII,55	9188
N2	15,2	obd., nobd., nbair./ofrk.	P	Nürnberg, Cent.VI,46f	7498
N3	15,1	obd., nobd., nbair./ofrk.	P	Nürnberg, Cent.VI,86	4274
N4	15	obd., wobd./oobd., alem./bair.	P	Nürnberg, Hs.23212	8625
O	14,2	ndd., ofäl.	V	Oldenburg, CimI74	7320
s	14,2	ndd., nndd., östl. nndd.	P	Schwerin, oSignatur	664
sa	15,2	obd., wobd., hchalem.	P	Sarnen, Cod.membr.33	8759
Sa	15,2	obd., wobd., hchalem.	P	Sarnen, Cod.chart.125	8653
Sb	15,1	obd., oobd., mbair.	P	Salzburg, Cod.23A22	7438
SG	16,1	obd., wobd., hchalem.	P	St.Gallen, Cod.Sang.1006 SDG	7926
sl*	15,2	obd., oobd.	P	St.Leonhard, oSignatur	177
SP	15	ndd., nndd.	V	St.Petersburg, Fond955op2Nr51	7093
St	15	md., wmd., rhfrk.-hess.	P	Strassburg, Ms.2267	7400
St2	14	obd., wobd., alem.	P	Strassburg, Cod.A100	8871
Stu	15,1	obd., wobd., schwäb.	P	Stuttgart, Cod.bibl.2°35	8687
T	14	obd., oobd., mbair.	P	Troppau, RA-6	9822
W	15,1	obd., oobd., mbair.	P	Wien, Cod.2969	8601
We	15,2	obd., oobd., nbair.	P	Weimar, Cod.Oct.4	6651
Wo	14,1	ndd., ofäl., südl. ofäl.	P	Wolfenbüttel, Cod. Guelf. 1082	4884

Tab. 1: Hss. des Anselm-Korpus und ihre manuell vorgenommenen Zuordnungen

DRUCKE

Sigle	Dat.	Sprachraum	Prosa/ Vers	Druckort u. Drucker	Größe
HA1521	16,1	ndd., nndd.	V	Lübeck, Arndes	7098
KÄ1492	15,2	md., wmd., rip.	V	Köln, Koelhoff d.Ä.	7591
KJ1499	15,2	md., wmd., rip.	V	Köln, Koelhoff d.J.	7537
Kr1522*	16,1	md., wmd., rip.	V		
N1500	16,1	md., wmd., rip.	V	Köln, Neuss	8124
N1509	16,1	md., wmd., rip.	V	Köln, Neuss	7913
N1514	16,1	md., wmd., rip.	V	Köln, Neuss	7669
N1514b	16,1	md., wmd., rip.	V		
s1495	15,2	obd., oschwäb.	P	Augsburg, Schaur	1874+444
s1496/7	15,2	obd., oschwäb.	P	Augsburg, Schaur	6214
StA1495	15,2	ndd., nndd.	V	Lübeck, Arndes	6694

Tab. 2: Drucke des Anselm-Korpus und ihre manuell vorgenommenen Zuordnungen

2. Automatische Verfahren

Die Textzeugen sollen nun automatisch geclustert, d.h. nach Ähnlichkeit gruppiert werden. Im ersten Schritt müssen dazu die Merkmale, die den Gruppen zugrunde liegen sollen, bestimmt werden. Es sollen dabei ausschließlich vollautomatische Verfahren zum Einsatz kommen. Linguistisches Wissen fließt hierbei auf folgende Art in die Verfahren ein: zum einen bei der Auswahl geeigneter Merkmale, zum andern bei der Wahl geeigneter Algorithmen und Maße zur Berechnung der Ähnlichkeiten.

2.1 Ähnlichkeitsmerkmale

Die Sprachraumgrenzen im deutschen Dialektraum beruhen vorwiegend auf Unterschieden im Lautstand. Beispielsweise zeigt die Karte in Abbildung 5[7] die Grenzen verschiedener deutscher Sprachräume und markiert die Isoglossen z.T. mit entsprechenden Varianten.

[7] Die Karte wurde entnommen von http://www.ruhr-uni-bochum.de/wegera/texte_1.htm (13.08.2014). Sie stammt ursprünglich aus Hermann Paul (Mittelhochdeutsche Grammatik. 25. Aufl., neu bearbeitet von Thomas Klein, Hans-Joachim Solms und Klaus-Peter Wegera. Mit einer Syntax von Ingeborg Schröbler, neu bearbeitet und erweitert von Heinz-Peter Prell. Tübingen 2007, S. 3) und wurde am Lehrstuhl Wegera mit den Ortspunkten der Textzeugen angereichert. Die Grau-Einfärbungen wurden von uns hinzugefügt und verdeutlichen die Grenzen der Sprachräume. Relevant für uns ist dabei nicht der genaue Verlauf einer Grenze, sondern die Frage, welche Sprachräume direkt angrenzen. Für die Bearbeitung der Karte wie auch für ein Programm zur Berechnung der Distanzen zwischen Regionen der Karte (s. Abschnitt 3.2) möchte ich Roman Klippert vielmals danken.

Z.B. unterscheidet die Variante *appel*/*apfel* den mitteldeutschen von dem oberdeutschen Sprachraum, s. die mit „e“ markierte Grenze im Kartenausschnitt in Abb. 1. Da sich solche Unterschiede oft auch in der geschriebenen Sprache widerspiegeln, liegt es auf der Hand, die Oberflächenform von Wörtern als Merkmal für eine Ähnlichkeitsberechnung zu nehmen.

Abb. 1: Kartenausschnitt mit der Grenze Rheinfränkisch – Alemannisch

Ein naheliegender Ansatz wäre hier, eine Reihe von geeigneten Lemmata (wie z.B. *Apfel*) auszuwählen und deren Schreibungen in den verschiedenen Textzeugen zu vergleichen, wie es in der historisch-vergleichenden Sprachwissenschaft üblich ist. Voraussetzung dafür ist allerdings, dass die Transkripte entsprechend lemmatisiert werden. Dies ist momentan noch nicht der Fall, zudem gibt es aktuell keine Tools, die historische Sprachdaten automatisch lemmatisieren könnten. Außerdem handelt es sich bei den Textzeugen um eher kurze Dokumente, so dass es wohl kaum geeignete Lemmata in ausreichend vielen Dokumenten geben würde.

Ein anderer Ansatz wurde in einer Pilotstudie von Anke Lüdeling[8] vorgeschlagen. Hier wurden zunächst fünf Versionen des „Vater unser“ aus verschiedenen Sprachstufen des Deutschen phonetisch und syntaktisch analysiert. Die Versionen wurden außerdem wort- und satzweise miteinander aligniert, d.h. sich entsprechende Wörter und Sätze wurden einander zugeordnet, so dass paarweise Differenzen zwischen phonetischen bzw. syntaktischen Eigenschaften berechnet werden konnten.

Ein solcher Ansatz setzt voraus, dass die zu vergleichenden Dokumente hinreichend ähnlich sind, so dass sie sinnvoll miteinander aligniert werden können. Die Transkripte im Anselm-Korpus sind wohl prinzipiell geeignet für Alignierungen, allerdings müssten sie manuell aligniert werden, da aktuell noch keine entsprechenden Alignierungsprogramme existieren.

Ich verfolge daher einen alternativen Ansatz, der ohne Alignierungen auskommt und statt einer vordefinierten Auswahl von Lemmata sämtliche Wortformen eines Dokuments einbezieht und die Wortformen dazu in Fragmente zerlegt. Beide Maßnahmen wirken sich positiv auf den Vergleich kleiner Datenmengen aus, da zum einen die Dokumente in ihrer Gesamtheit berücksichtigt werden, zum andern durch die Zerlegung in Fragmente die Zahl der Datenpunkte erhöht wird. Diesen Ansatz werde ich in verschiedenen Variationen durchführen, die in den folgenden Abschnitten beschrieben werden.[9]

---

8 Anke Lüdeling: Using corpora in the classification of language relationships, in: Zeitschrift für Anglistik und Amerikanistik. Special Issue on ‚The Scope and Limits of Corpus Linguistics‘, 2006, S. 217–227.

9 In einer Pilotstudie (Stefanie Dipper, Bettina Schrader: Computing Distance and Relatedness of Medieval Text Variants from German, in: Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing, hg. v. Angelika Storrer u.a., Berlin 2008, S. 39–51.) habe ich bereits ähnliche Verfahren auf einem winzigen Datenfragment mit manuell alignierten Daten durchge-

2.1.1 Sequenzen von Buchstaben (diplomatisch)

In der ersten Variation dienen die diplomatischen Transkripte als Textgrundlage. Zunächst wird jede diplomatische Wortform in Fragmente oder N-Gramme (d.h. Folgen von *n* Zeichen) von 1–3 Buchstaben zerlegt. Beispielsweise werden die beiden Varianten der Sequenz „zu den Juden“ in (3) zerlegt in die in (4) bzw. (5) gelisteten N-Gramme. „#“ und „##“ kennzeichnen den Wortanfang bzw. das Wortende. In (a) stehen Unigramme, d.h. 1-buchstabige Fragmente, in (b) Bigramme und in (c) Trigramme.

(3) a. *zů den Iuden* (Stu)
b. *czu den iuden* (B)

(4) a. z - ů - d - e - n - I - u - d - e – n
b. #z - zů - ů # - #d - de - en - n# - #I - Iu - ud - de - en - n#
c. ##z - #zů - zů# - ů## - ##d - #de - den - en# - n## - ##I - #Iu - Iud - ude - den - en# - n##

(5) a. c - z - u - d - e - n - i - u - d - e - n
b. #c - cz - zu - u# - #d - de - en - n# - #i - iu - ud - de - en - n#
c. ##c - #cz - czu - zu# - u## - ##d - #de - den - en# - n## - ##i - #iu - iud - ude - den - en# - n##

Für jedes Dokument wird nun ein ‚Profil‘ berechnet, das sich aus den im Dokument vorkommenden Buchstabensequenzen und ihren Frequenzen ergibt. Die Abfolge der einzelnen Sequenzen im Dokument ist dabei irrelevant. Die Profile der Dokumente können anschließend miteinander verglichen werden: Je mehr N-Gramme sie miteinander teilen, desto ähnlicher sind sich die Dokumente. Beispielsweise sind in (4a) und (5a) 2x8 = 16 von insgesamt 21 Buchstaben identisch, so dass sich eine Übereinstimmung von 76% ergibt, s. Tab. 3. In (4b)/(5b) sind 16 von 27 Bigrammen, d.h. 59% identisch, in (4c)/(5c) sind es 18 von 33 Trigrammen, d.h. 55%. Je länger die N-Gramme, desto spezifischer werden die Vergleiche und desto geringer die Übereinstimmung. Im Extremfall werden komplette Wörter miteinander verglichen.

Was der Vergleich hier noch nicht berücksichtigt, ist der Einfluss der Textlänge: Vergleicht man einen sehr kurzen mit einem sehr langen Text, kann natürlich nur ein Bruchteil der N-Gramme übereinstimmen. D.h. die Frequenzen der N-Gramme müssen zunächst normalisiert werden, indem die Frequenzen der einzelnen N-Gramme relativ zur Gesamtanzahl von N-Grammen berechnet werden. Unter 2.2 wird ein entsprechend modifiziertes Ähnlichkeitsmaß eingeführt.

---

führt. Die hier berichteten Ergebnisse beruhen auf einer unvergleichlich größeren Datenmenge und nutzen ausschließlich automatische Verfahren.

Unigramm	Stu	B	identisch
C	0	1	0
D	2	2	4
E	2	2	4
I	1	0	0
I	0	1	0
N	2	2	4
U	1	2	2
Ů	1	0	0
Z	1	1	2
Summe	10	11	16/21

Tab. 3: Vergleich von Unigramm-Frequenzen in zwei Dokumenten

2.1.2 Sequenzen von Buchstaben (simplifiziert)

Die verschiedenen N-Gramm-Vergleiche im Abschnitt 2.1.1 ergaben Übereinstimmungen zwischen 76% und 55%. Vergleicht man diese Ergebnisse mit den eigenen Intuitionen, so würde man vermutlich einen höheren Grad an Übereinstimmung zwischen den Sequenzen (3a) und (3b) erwarten – tatsächlich sind sie ja fast identisch.

Dieser Intuition soll in der zweiten Variation Rechnung getragen werden. Die diplomatischen Zeichen werden dazu simplifiziert. Konkret werden alle Groß- durch Kleinbuchstaben, Superskripte durch entsprechende nachgestellte Buchstaben und das Schaft-s durch normales *s* ersetzt. Außerdem werden alle Diakritika gelöscht und die Wortgrenzen modernen Konventionen angepasst. Die beiden Sequenzen aus (3) sehen dann wie in (6) aus (die diplomatischen Ausgangsformen sind jeweils mit angegeben). Die simplifizierte komplette Version des Beispieltexts in (1) wird in Abb. 2 gezeigt.

(6) a. *zů den Iuden* → *zuo den iuden* (Stu)
b. *czu den iuden* → *czu den iuden* (B)

Damit fallen beispielsweise die Unterschiede zwischen den Unigrammen *ů* (Stu) und *u* und *o* (B) und zwischen *I* (Stu) und *i* (B) weg. Die prozentuale Übereinstimmung von (6a) und (6b) steigt dann auf 91% bei Unigrammen (20 von 22 Buchstaben sind identisch), 79% bei Bigrammen und 71% bei Trigrammen und liegt damit wesentlich höher als bei den diplomatischen Vergleichen.

Es sind natürlich weitere Simplifizierungen vorstellbar, z.B. die Gleichsetzung von *v* und *u* oder von *i* und *j*. Ich beschränke mich hier bewusst auf die einfachsten Ersetzungen, die keine weiteren historisch-linguistischen Kenntnisse erfordern.

2.1.3 Sequenzen von Phonen

Die dritte Variation vergleicht phonetische Repräsentationen der Überlieferungen. Die zugrunde liegende Idee ist, dass auf diese Art bestimmte graphematische Unterschiede zusammenfallen, so z.B. die Grapheme *v* und *f*. (7) zeigt ein entsprechendes Beispiel aus den Handschriften Stu und B, in simplifizierter Schreibung und in der daraus automatisch abgeleiteten phonetischen Repräsentation (in SAMPA-Notation, s.u.).

(7) a. *frow* → fRo: (Stu)
b. *vrouwe* → fRu:vE (B)

Für die Erzeugung der phonetischen Repräsentation wird das Programm txt2pho genutzt, das Teil des Sprachsynthesesystems MBROLA ist.[10] Die phonetische Repräsentation wird in SAMPA-Notation wiedergegeben. SAMPA ist ein sprachabhängiges phonologisch-phonetisches Alphabet, das ausschließlich ASCII-Zeichen nutzt und sich daher besonders für die automatische Verarbeitung eignet. In der SAMPA-Version für die deutsche Sprache kodiert z.B. *R* den uvularen Vibranten (wie in *Frau*), @ den Schwa-Laut (wie in *bitte*) und 6 den Schwa-Laut mit velarer Färbung (wie in *mir*). Der Doppelpunkt markiert lange Vokale. Kleingeschriebene Vokale sind gespannt, großgeschriebene ungespannt (bis auf *E:*).[11] txt2pho wurde für modernes Deutsch entwickelt und enthält u.a. ein Lexikon mit Ausspracheinformationen. Dem Lexikon unbekannte Wörter werden mit Hilfe von Regeln abgebildet. Viele der Wortformen in den Anselm-Textzeugen sind solche unbekannten Wörter.

Unsere beiden Beispielsequenzen für den N-Gramm-Vergleich sind in (8) dargestellt (auch hier ist die simplifizierte Ausgangsform jeweils mit angegeben). Es ist offensichtlich, dass die automatische Abbildung nicht fehlerfrei ist. Beispielsweise wird das Graphem *z* hier auf stimmloses *s* abgebildet (8a), *cz* auf *tS* (gesprochen [tʃ]) (8b). Auf der phonetischen Ebene gibt es in diesem Beispiel dadurch weniger Übereinstimmung als auf der simplifizierten.

(8) a. *zuo den iuden* → sUo: de:n Iu:dEn (Stu)
b. *czu den iuden* → tSU de:n Iu:dEn (B)

Die einfachste Berechnungsmethode ist, den Doppelpunkt als separates Zeichen zu behandeln. Damit ergeben sich folgende prozentualen Übereinstimmungen: Unigramme 81% (22 von 27 Zeichen sind identisch); Bigramme 73%; Trigramme 72%.

Die automatisch erzeugte SAMPA-Repräsentation des kompletten Beispieltexts aus (1) wird in Abbildung 2 gezeigt. Dort findet sich ein weiterer typischer Fehler: *vnd* abgebildet auf *faUEnde:*. Die drei Buchstaben werden also einzeln buchstabiert,

---

[10] *txt2pho* ist frei verfügbar unter http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/txt2pho.

[11] Eine Übersicht über das deutsche SAMPA-Alphabet findet sich unter http://coral.lili.uni-bielefeld.de/Documents/sampa-d-vmlex.html.

da es sich (scheinbar) um eine Sequenz ohne Vokal handelt. Solche und andere Fehler kommen in allen Textzeugen vor und können die Ergebnisse verfälschen, v.a. wenn es sich um systematische Fehler handelt, die häufig auftreten. Ist aber die Datenmenge groß genug, gehen einzelne Fehler im sogenannten ‚Rauschen' unter und können statistisch ignoriert werden – was ich in diesem Artikel auch tun werde.

2.1.4 Sequenzen von Wortarten

Die bisherigen Variationen betrafen alle den Lautstand der verschiedenen Textzeugen, indem oberflächennahe Merkmale verglichen wurden. In der vierten und letzten Variation hingegen werden die Textzeugen anhand syntaktischer Eigenschaften miteinander verglichen. Dazu werden die simplifizierten Wortformen zunächst automatisch normalisiert, d.h. auf entsprechende moderne Wortformen (in Kleinschreibung) abgebildet und anschließend automatisch getaggt, d.h. nach ihrer Wortart kategorisiert. Für die Normalisierung wird das Programm Norma genutzt,[12] für das Wortarten-Tagging der RFTagger.[13]

Die hier verwendeten Wortart-Kategorien folgen dem STTS-Standard.[14] Kodiert werden neben der Wortart auch Flexionseigenschaften. Beispielsweise werden finite Vollverben der Kategorie „VVFIN" (Verb, voll, finit) zugeordnet, Infinitive von Auxiliaren der Kategorie „VAINF" (Verb, Auxiliar, infinitiv) etc. Anhand der

---

12 Marcel Bollmann: (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool, in: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), hg. v. Francesco Mambrini, Marco Passarotti, Caroline Sporleder, Lisbon 2012, S. 3–14.

13 Helmut Schmid, Florian Laws: Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, in: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), Manchester 2008, S. 777–784.

Die hier verwendete Version von Norma wurde auf rund 50.000 manuell normalisierten Wortformen aus dem Anselm-Korpus trainiert. Der hier verwendete RFTagger wurde auf dem TIGER-Korpus (Sabine Brants u.a.: TIGER: Linguistic Interpretation of a German Corpus, in: Research on Language and Computation, Special Issue, 2(4), hg. v. Erhard Hinrichs, Kiril Simov, 2004, S. 597–620.) und der TüBa-D/Z (Heike Telljohann u.a.: Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen, Januar 2012; www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf) ohne Satzzeichen und in Kleinschreibung trainiert (s. Marcel Bollmann: Automatic Normalization for Linguistic Annotation of Historical Language Data, Bochumer Linguistische Arbeitsberichte 13, Bochum 2013).

14 Anne Schiller u.a.: Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen, 1999 (www.sfs.uni-tuebingen.de/resources/stts-1999.pdf). Eine Übersicht über die Kategorien des STTS gibt es unter http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html.

Wortartsequenzen kann dann z.B. zwischen Sprachen unterschieden werden mit den Abfolgen Vollverb > Auxiliar vs. Auxiliar > Vollverb.

Unsere Beispielsequenzen in normalisierter und getaggter Form werden in (9) gezeigt, zusammen mit der simplifizierten Ausgangsform. APPR markiert Präpositionen („Adposition, Prä-Stellung"), ART Artikel und NN normale Nomen.

(9) a. zuo den iuden → zu den juden → APPR ART NN (Stu)
b. czu den iuden → zu den juden → APPR ART NN (B)

Die Ergebnisse des automatischen Normalisierers sind nicht immer so gut wie im gezeigten Beispiel. In der normalisierten Form von Beispiel (1) (s. Abbildung 2) findet sich ein typischer Fehler: die Verwechslung von *dass* und *das*. Da der verwendete Normalisierer nur jeweils Einzelwörter verarbeitet, kann keine Kontextinformation zur Disambiguierung genutzt werden. (10a) zeigt eine fehlerhaft normalisierte Sequenz, mit den entsprechenden Konsequenzen für die Wortart-Zuweisung. *dass* kommt zweimal vor und wird als KOUS (subordinierende Konjunktion) analysiert. Korrekt wäre ART bzw. PRELS (Relativpronomen), s. (10b). Außerdem wird das unflektierte *jüngst* als ADV (Adverb) analysiert statt als ADJA (attributives Adjektiv). Solche Fehler kommen in allen Textzeugen vor und können sich daher unter Umständen sogar gegenseitig aufheben.

(10) a. dass/KOUS jüngst/ADV mahl/NN dass/KOUS er/PPER nahm/VVFIN
(Stu, automatisch)
b. das/ART jüngste/ADJA Mahl/NN, das/PRELS er/PPER nahm/VVFIN
(Stu, manuell)

Beispiel (11) zeigt den Anfang einer niederdeutschen Vers-Handschrift (Kh), für den der Normalisierer sehr schlechte Ergebnisse liefert (der Grund dafür ist, dass in den Trainingsdaten von Norma keine niederdeutschen Daten enthalten waren, vgl. Anm. 11). (11a) zeigt die diplomatische Version, (11b) die phonetische Version und (11c) die normalisierte mit Wortart.

(11) a. *Anſylmus was eyn hillich man*
he hadde dar lange na geſtan
Dat he gerne wolde weten
wat vnſe here hadde beſeten (Kh, 232r,1–4)
„Anselmus war ein heiliger Mann, / er hatte lange danach gestrebt, /
dass er gerne wissen wollte / was unserem Herrn widerfahren war."
b. anzYlmUs vas aIn hIlIC man
he: hadE da:6 laN@ na: g@Sta:n
da:t he: gERn@ vOldE v@tn
va:t fnzE he:R@ hadE b@sEtn

c. anselmus/NE war/VAFIN ein/ART hellig/ADJD man/FM
he/FM hande/FM dar/NE lange/ADV nach/APPR gesan/NE
das/ART he/FM gerne/ADV wollte/VMFIN weden/NN
wad/NE unsere/PPOSAT herr/NN hande/ADV besten/ADJD

Es ist offensichtlich, dass auf Basis dieser Daten keine Aussagen über syntaktische Eigenschaften des Niederdeutschen gemacht werden können. Allerdings könnten sich die Daten dennoch für eine vergleichende Auswertung eignen. Beispielsweise sind in der niederdeutschen Handschrift Kh überproportional viele Wortformen mit FM (Fremdwort) getaggt, nämlich 9% (641 von 7138). In der oberdeutschen Handschrift Stu sind es hingegen unter 1% (65 von 8687). D.h. wir können u.U. auch Fehler der automatischen Verarbeitung als Indiz für eine sprachräumliche Zugehörigkeit nutzen – was ich im Folgenden tun werde, indem Fehler wie schon oben bei der phonetischen Repräsentation generell ignoriert werden, d.h. unverändert mit in die statistischen Berechnungen eingehen.

Diplomatisch: Sant aunſhalm waz von herczen fro ſin frag hů̊b er an vnd ſprach zů̊ ir ſag mir liebů̇ frow wie was der anfank der marter dines lieben kindes vnſer frow ſprach do min kint het geſſen mit ſinen iungern vor ſiner marter daz Iungſt maſſ̄ daz er nam vnd do ſi von dem tiſch vff ſtů̊nden do gieng iudaz ſcariot zů̊ den Iuden der fů̊rſten vnd kam ainſ̄ gedinges mit in v̊̇ber ain daz er min kint wolt verrå̊tten

Simplifiziert: sant aunshalm waz von herczen fro sin frag huob er an vnd sprach zuo ir sag mir liebui frow wie was der anfank der marter dines lieben kindes vnser frow sprach do min kint het gessen mit sinen iungern vor siner marter daz iungst mass daz er nam vnd do si von dem tisch vffstuonden do gieng iudaz scariot zuo den iuden der fuersten vnd kam ains gedinges mit in viberain daz er min kint wolt verraetten

Phonetisch: zant aUnshalm vats fOn he:6tSEn fRo: zIn fRa:k hu:Op e:6 an faUEnde: SpRa:x tsUo: IR za:k mi:6 li:pu:i: fRo: vi: vas de:6 anfaNk de:6 maRt6 di:nEs li:bn kInd@s fnz6 fRo: SpRa:x do: mIn kInt ha@t gEsn mIt zi:n@n IUng6n fo:6 zi:n6 maRt6 dats IUNst ma:s dats e:6 na:m faUEnde: do: zi: fOn de:m tIS fStu:o:ndn do: gi:ng aIu:dats kaRi:o:t sUo: de:n Iu:dEn de:6 fYRstn faUEnde: ka:m aIns g@dIN@s mIt In vi:b@RaIn dats e:6 mIn kInt vOlt fEREtn

Normalisiert mit Wortart: sankt/NE anselm/NE war/VAFIN von/APPR herzen/NN froh/ADJD sein/PPOSAT frage/NN hob/VVFIN er/PPER an/PTKVZ und/KON sprach/VVFIN zu/APPR ihr/PPER sag/VVIMP mir/PPER liebe/VVFIN frau/NN wie/PWAV war/VAFIN der/ART anfang/NN der/ART marter/NE deines/PPOSAT lieben/ADJA kindes/NN unser/PPOSAT frau/NN sprach/VVFIN da/ADV mein/PPOSAT kind/NN hat/VAFIN gegessen/VVPP mit/APPR seinen/PPOSAT jüngern/NN vor/APPR seiner/PPOSAT marter/NE dass/KOUS jüngst/ADV mahl/NN dass/KOUS er/PPER nahm/VVFIN und/KON da/KOUS sie/PPER von/APPR dem/ART tisch/NN aufstanden/VVPP da/ADV ging/VVFIN iudas/NE scharioth/NE zu/APPR den/ART juden/NN der/ART fürsten/NN und/KON kam/VVFIN eines/PIS dinges/NE mit/APPR ihn/PPER überein/PTKVZ dass/KOUS er/PPER mein/PPOSAT kind/NN wollte/VMFIN verraten/VVFIN

Abb. 2: Ein Beispielfragment in verschiedenen Versionen (Stu, 96va,21–31)

2.2 Kosinus-Ähnlichkeit

Im Abschnitt 2.1 wurden vier Variationen von Repräsentationsformaten vorgestellt, in denen die verschiedenen Textzeugen verglichen werden sollen. Ein zweiter Variationsparameter ist das Maß, das für den Vergleich genommen wird. In den Beispielberechnungen in 2.1 habe ich ein sehr simples Maß angewendet, bei dem identische N-Gramme gezählt werden. Ein schon erwähntes Problem dieses Maßes ist, dass es die potenziell stark variierende Länge der zu vergleichenden Dokumente nicht berücksichtigt.

Das Maß, das ich daher für die Ähnlichkeitsberechnung anwende, heißt Kosinusähnlichkeit. Zunächst wird jedes Dokument als ein Vektor dargestellt. Die Dimensionen des Raums werden durch alle vorkommenden N-Gramme im Korpus bestimmt. Die Koordinaten des Vektors bestimmen sich durch die Frequenzen der N-Gramme. Kosinus-Ähnlichkeit berechnet dann den Winkel zwischen den Vektoren. Zeigen die Vektoren in eine ähnliche Richtung, ist der Winkel klein und die Dokumente sind sehr ähnlich zueinander. Zeigen sie in unterschiedliche Richtungen, wächst der Winkel an.

Im Falle von maximal drei Dimensionen kann das Maß auch visuell veranschaulicht werden. Nehmen wir an, wir hätten ein Korpus aus fünf extrem kurzen Dokumenten D1–D5, wie in (12a–e).

(12) a. D1: *zu* → (0,1,1)
b. D2: *c* → (1,0,0)
c. D3: *czu* → (1,1,1)
d. D4: *zuu* → (0,2,1)
e. D5: *zuzu* → (0,2,2)

Es kommen insgesamt drei verschiedene Buchstaben vor: *c*, *u*, *z*. Sie definieren die drei Dimensionen des Vektorraums, *c* die horizontale Achse, *u* die vertikale und *z* die Tiefenachse, s. Abbildung 3.

Dokument D1 wird dann als Vektor (0,1,1) dargestellt, da es 0x den Buchstaben *c* enthält und jeweils 1x die beiden anderen Buchstaben. Im Raum zeigt der Vektor schräg nach hinten oben (s. die Markierung mit „D1“ in Abb. 3). D2 stellt gewissermaßen das Komplement zu D1 dar: Es enthält nur 1x *c* und sonst keinen weiteren Buchstaben. D1 und D2 sind sich also maximal unähnlich. Daher stehen die beiden Vektoren im rechten Winkel zueinander, und ebenso D2 und D4/D5. Im hier genutzten Maß der Kosinusähnlichkeit entspricht dieser Winkel dem Wert 0.

D1 und D3 teilen sich immerhin zwei Buchstaben, *u* und *z*. Daher ist der Winkel zwischen ihnen (in Abb. 3 mit α gekennzeichnet) kleiner als 90° und der Kosinus-Wert ist 0,82. D4 wiederum ist recht ähnlich zu D1, enthält aber 2x *u*, so dass D4’s Steigungswinkel doppelt so groß ist wie der von D1 (Kosinuswert: 0,95). Der Vektor von D5 schließlich ist identisch mit dem von D1 (Kosinuswert: 1). Durch die Übersetzung von Frequenzen in Koordinaten wird also auch das Problem unterschiedlich langer Dokumente gelöst.

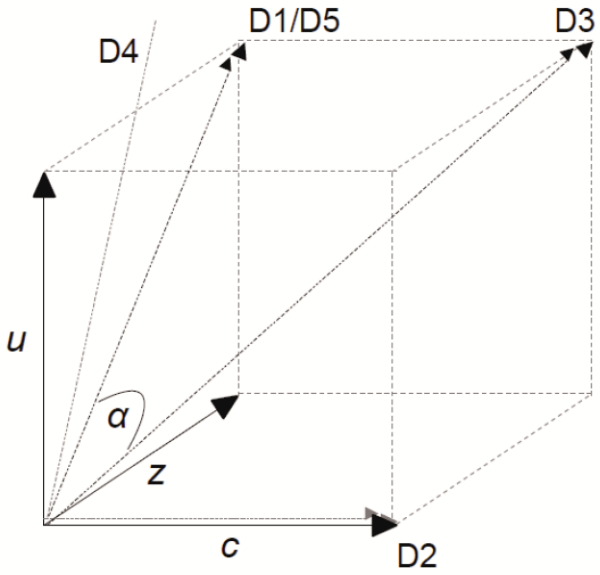

Abb. 3: Dreidimensionaler Vektorraum mit fünf Dokumenten

Tab. 4 zeigt die Werte der Kosinusähnlichkeiten paarweise zwischen allen fünf Dokumenten. Beispielhaft soll hier der Wert für das Dokumentenpaar D1–D3 berechnet werden, s. (13).

(13) a. $\text{sim}(a,b) = \frac{a*b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n}(a_i*b_i)}{\sqrt{\sum_{i=1}^{n}(a_i)^2}*\sqrt{\sum_{i=1}^{n}(b_i)^2}}$

b. $\text{sim}((0,1,1),(1,1,1)) = \frac{(0*1)+(1*1)+(1*1)}{\sqrt{0+1+1}*\sqrt{1+1+1}} = \frac{2}{2.45} = 0.82$

	D1	D2	D3	D4	D5
D1	100	0	82	95	100
D2		100	58	0	0
D3			100	77	82
D4				100	95
D5					100

Tab. 4: Kosinus-Ähnlichkeiten der fünf Dokumente (in Prozenten)

Diese Methode, Frequenzprofile verschiedener Dokumente für Ähnlichkeitsberechnungen zu nutzen, kann auf Frequenzen beliebiger Merkmale angewendet werden, wie z.B. sämtliche in 2.1 vorgestellten Merkmale – was wir weiter unten auch tun werden.

2.3 Clustering

In Abschnitt 2.2 haben wir gesehen, wie Frequenzprofile dazu genutzt werden können, mithilfe des Kosinusmaßes paarweise Ähnlichkeiten zwischen Dokumenten zu berechnen. Was jetzt noch fehlt, ist eine Kombination der paarweisen Einzelwerte, so dass jeweils Gruppen zueinander ähnlicher Dokumente gebildet werden können, sogenannte Cluster.

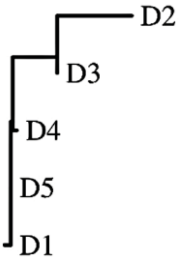

Abb. 4: Visualisierung der Ähnlichkeiten

Hierbei setze ich das Programm phylip ein, das verschiedene Clustering-Algorithmen anbietet.[15] Der Algorithmus Neighbor Joining, den ich hier nutze, vereint zunächst die beiden ähnlichsten Dokumente zu einem Cluster und berechnet für diesen Cluster einen neuen Ähnlichkeitswert aus dem Durchschnitt der beiden darin enthaltenen Dokumente. Dann wird der Algorithmus auf die restlichen Dokumente und den bereits bestehenden Cluster angewendet und wiederum die zwei ähnlichsten miteinander vereint. Der Vorgang wird solange wiederholt, bis alle Dokumente Teil eines Clusters sind.

Das Ergebnis des Clusterings kann in Form eines Baumes visualisiert werden, s. Abb. 4. Die Wurzel wird dabei von den zuerst geclusterten Dokumenten gebildet. Die Differenz zwischen den Dokumenten wird durch die horizontale Distanz repräsentiert.

3. Auswertung

3.1 Paarweise Distanzen

Ich beginne die Auswertung mit einigen Beobachtungen zu den automatisch berechneten paarweisen Distanzen.

- Generell gilt, dass bei den 3x3 = 9 buchstabenbasierten Vergleichen (diplomatisch, simplifiziert und phonetisch, jeweils mit Uni-, Bi- und Trigrammen) stets ripuarische Drucke auf den vordersten drei Plätzen landen, d.h. die ähnlichsten Paarungen bilden, und zwar mit Kosinuswerten von quasi 100%. Dabei steht entweder das Paar KÄ1492–KJ1499 (in 6 von 9 Variationen) oder das Paar N1509–N1514 (in 3 Variationen) ganz oben.
- Im Kontrast dazu stehen bei den drei wortartbasierten Vergleichen mit Hk, M2 und M4 drei bairische Handschriften ganz oben (mit Werten von >99% bei den Uni- und Bigrammen und 97% bei den Trigrammen).

[15] *phylip* ist frei erhältlich unter http://evolution.genetics.washington.edu/phylip.html (s. Joseph Felsenstein: PHYLIP – Phylogeny Inference Package (Version 3.2), in: Cladistics 5, 1989, S. 164–166).

– Am unteren Rand finden sich bei den buchstabenbasierten Vergleichen die extrem kurzen Fragmente wieder: au (145 Wortformen), hb (55) und b2 (746). Die Kosinuswerte der jeweils unähnlichsten Paarungen bewegen sich dabei grob zwischen 89% (Unigramme) und 65% (Trigramme); die Werte für diplomatische Formen liegen noch darunter: zwischen 85% und 50%.

  Bei den wortartbasierten Vergleichen kommt außerdem die umfangreichste Handschrift Ka hinzu (14.842 Wortformen). Hier liegen die minimalen Kosinuswerte bei 66% (Unigramme), 26% (Bigramme) bzw. 4% (Trigramme).

– Lässt man nun aus statistischen Gründen alle Textzeugen außer Acht, die kürzer als 4.000 Wortformen sind (das betrifft neun Dokumente), so ändert sich nichts an den übrigen Paarungen – die Winkel zwischen den verbleibenden Dokumenten bleiben davon ja unberührt. Die neuen Schlusslichter sind dann v.a. Paarungen zwischen niederdeutschen und mittelbairischen Textzeugen: u.a. die Paare StA1495–Sb, StA1495–Me, StA1495–M7 und Kh–W, Wo–W auf der Buchstaben-Ebene, und die Paare SP–Ka, D2–Ka, Kh–Ka auf der Wortartebene. Die meisten der niederdeutschen Textzeugen sind dabei allerdings Vers-Fassungen (bis auf Wo).

– Möchte man den Einfluss des Parameters ‚Prosa/Vers' ausschließen, so kann man die Untersuchung auf reine Prosa-Paarungen beschränken. Die sich dann neu ergebenden Schlusslichter bestehen wieder größtenteils aus niederdeutsch–mittelbairischen Paarungen (auf der Buchstabenebene): z.B. Wo–W, Wo–M9; aber es sind auch mittelbairische und alemannische Kombinationen mit der thüringischen Handschrift B vertreten, z.B. B–W, B–N4, B–Ka.[16]
  Auf der Wortart-Ebene finden sich v.a. die Paarungen Wo–M und Wo–Ka am Schluss, eine niederdeutsch–mittelbairische bzw. eine niederdeutsch–alemannische Kombination.

Diese Ergebnisse sehen also vielversprechend aus: Die sich ähnlichsten Dokumente kommen jeweils aus dem gleichen Sprachraum: ripuarisch bei den Buchstaben-Vergleichen, bairisch bei den Wortartvergleichen. Und die unähnlichsten Dokumente kommen aus weit entfernten Sprachräumen wie niederdeutsch–bairisch.

Im Folgenden gehen wir einen Schritt weiter und betrachten die Gesamtheit aller Paarungen, d.h. die sich ergebenden Cluster. Dazu soll das automatische Clustering aus 2.3 mit der manuell vorgenommenen sprachräumlichen Zuordnung verglichen werden. Dazu müssen allerdings zunächst Ähnlichkeitswerte und Cluster aus den manuellen Zuordnungen berechnet werden. Im anschließenden Schritt werden die Cluster der manuellen und der automatischen Zuordnungen miteinander verglichen.

[16] Dass die mbair. Textzeugen hier so prominent auftreten, liegt natürlich auch daran, dass sie die im Korpus am stärksten vertretene Sprachregion darstellen.

3.2 Ähnlichkeitswerte und Cluster für die manuellen Zuordnungen

Wir berechnen die Ähnlichkeitswerte auf Basis der Zuordnungen in den Tabellen 1 und 2. Mithilfe der Landkarte in Abb. 5 werden zunächst alle direkt benachbarten Sprachräume bestimmt. Dabei werden drei Stufen von Sprachräumen unterschieden: (i) die Einteilung in Ober-, Mittel- und Niederdeutsch; (ii) innerhalb dieser Einteilung größere Räume wie z.B. wobd. (westoberdeutsch), wmd. (westmitteldeutsch) oder ofäl. (ostfälisch); (iii) und schließlich Feinkategorisierungen wie z.B. schwäb. (schwäbisch), rip. (ripuarisch) oder elbofäl. (elbostfälisch). In der Karte sind Sprachräume der Stufe (ii) im gleichen Grauton wiedergegeben und feinere Unterteilungen der Stufe (iii) durch unterschiedliche Schraffierungen markiert.

In den Tabellen 1 und 2 sind die drei Stufen in der Spalte ‚Sprachraum' aufgeführt, durch Kommata voneinander abgetrennt. Nicht für alle Textzeugen sind alle drei Stufen angegeben (z.B. nicht für Kh, SP oder sl), und manche sind mischklassifiziert (z.B. D4 oder St). Für die Distanzberechnungen werden unterspezifizierte Dokumente (wie Kh, das nur nd., nnd. zugeordnet ist, also keinem Raum der Stufe (iii)) disjunktiv verteilt auf sämtliche Unterräume (Kh wird z.B. östl. nnd. und alternativ dem westlichen ‚Restraum' von nnd. (der in der Karte ohne eigene Benennung ist) zugeordnet). Bei Mischklassifikationen wird genauso verfahren: die Dokumente werden alternativ beiden Räumen zugeordnet (St wird z.B. alternativ rhrfk. bzw. hess. zugeordnet). Bei den Berechnungen werden zunächst getrennt Distanzen für alle Alternativen berechnet und davon der Durchschnittswert genommen.

Distanzen ergeben sich durch Überschreitungen von Sprachraumgrenzen. Die Überschreitung einer Grenze der Stufe (i) ‚kostet' 2, eine Grenze der Stufe (ii) 1 und eine Grenze der Stufe (iii) 0,5. Die Anzahl und Art der Grenzüberschreitungen, die man benötigt, um von einem Textzeugen zur anderen zu gelangen, kann anhand der Landkarte in Abb. 5 bestimmt werden. Bei jeder Grenzüberschreitung werden die Distanzen addiert. Beispielsweise liegt zwischen B2 (rhfrk.) und Stu (schwäb.) die Distanz 2, weil die Sprachgrenze zwischen Md. und Obd. überquert werden muss, eine Grenze der Stufe (i). Zwischen B2 (rhfrk.) und SG (hchalem.) liegt die Distanz bei 2,5, weil zusätzlich noch eine Grenze der Stufe (iii) überquert wird, entweder die zwischen Niederalemannisch und Hochalemannisch oder die zwischen Schwäbisch und Hochalemannisch. Dabei wird immer der kürzest mögliche Pfad genommen, d.h. der mit der geringsten Distanz.

Karte 1: *Interrogatio St. Anselmi*: Überlieferung

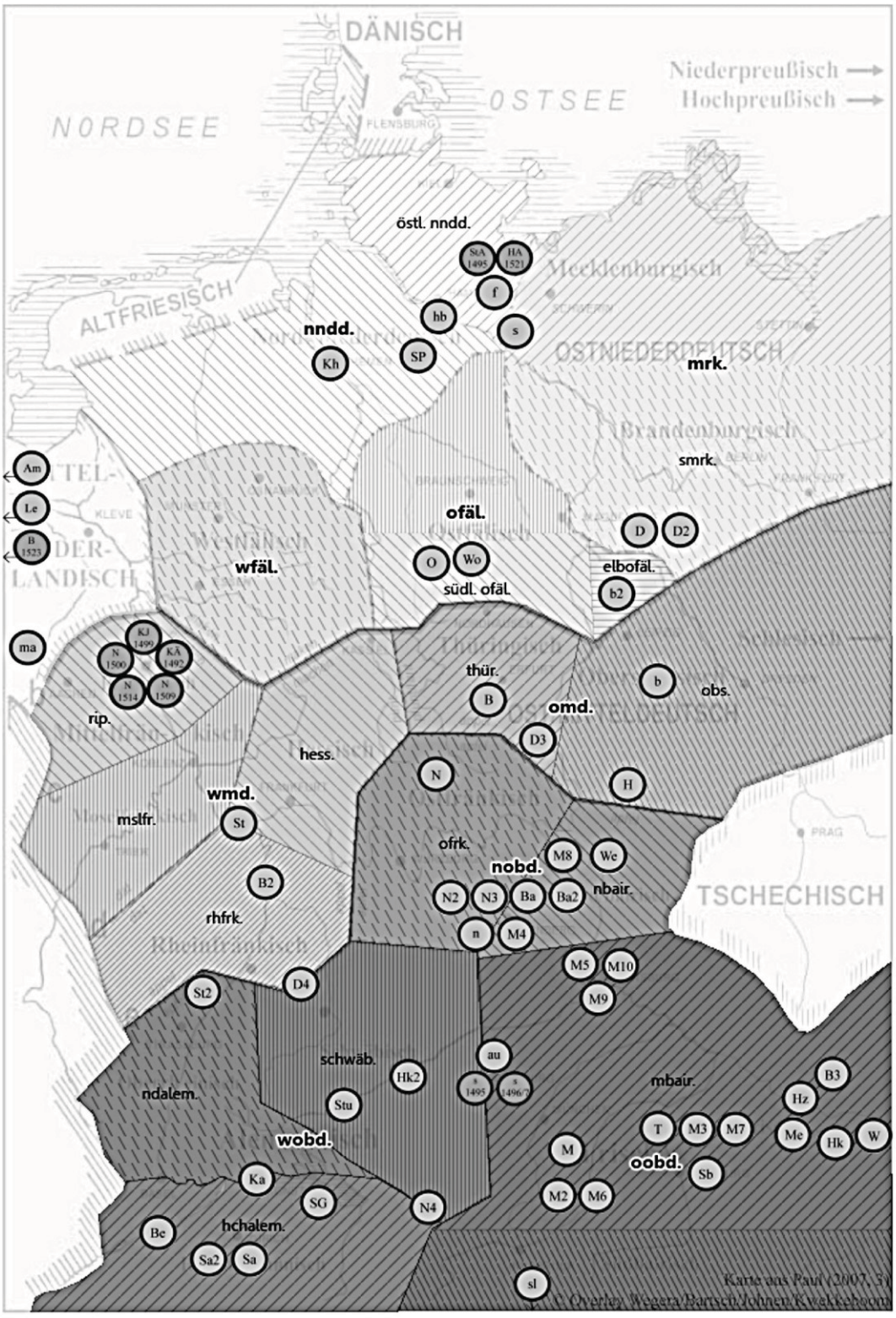

Abb. 5: Die Verteilung der Überlieferungsträger des Anselm-Textes innerhalb der deutschen Sprachraumgrenzen[17]

[17] Karte (modifiziert) aus Wegera (Anm. 1).

Anschließend werden die Distanzwerte normalisiert und in ein Ähnlichkeitsmaß konvertiert, das mit den automatisch bestimmten Kosinus-Werten verglichen werden kann. Dazu wird der maximal vorkommende Distanzwert (6,5)[18] auf den Wert 0 gesetzt, was dem minimal möglichen Kosinus-Ähnlichkeitswert entspricht. Dokumente desselben Sprachraums (z.B. Be und SG, beide hchalem.) haben einen Distanzwert von 0, da keine Grenze überschritten wird; das entspricht einem Ähnlichkeitswert von 1. Eine mittlere Distanz von 2,5 (z.B. zwischen B2 und SG) entspricht dem (gerundeten) Ähnlichkeitswert 0,615.

Die berechneten Ähnlichkeitswerte nutzen wir, wie schon in Abschnitt 2.3, um daraus Cluster von ähnlichen Dokumenten zu bilden. Die gerade berechneten Werte streuen allerdings deutlich mehr als die Ähnlichkeitswerte aus den automatischen Vergleichen – z.B. kommt bei den automatischen Vergleichen nie der Wert 0 („totale Unähnlichkeit“) vor. Da dieser „Fehler“ jedoch bei allen Vergleichen gleichermaßen involviert ist, gleicht er sich aus.

3.3 Vergleich der Cluster

Wie ähnlich sind sich nun die Cluster, die sich aus der manuellen Zuordnung und aus den automatischen Variationen ergeben? Und welche der automatischen Methoden trifft die manuelle Einteilung am besten? Dazu untersuchen wir zunächst einige Ergebnisse qualitativ und vergleichen anschließend die Cluster quantitativ mit Hilfe eines Ähnlichkeitsmaßes für Clustervergleiche. Ich beschränke mich bei den Clustervergleichen auf die vielversprechendsten Variationen: (i) simplifizierte Bi- und Trigramme, (ii) phonetische Bi- und Trigramme und (iii) Wortart-Bigramme.[19]

3.3.1 Qualitativer Vergleich

Ich beginne mit den buchstabenbasierten Clustern.

– Die großen Sprachräume der Stufe (i) (Ober-, Mittel- und Niederdeutsch) werden in allen buchstabenbasierten Clustern sehr gut wiedergegeben. Es finden sich nur zwei Ausreißer: Die alemannische Handschrift St2 wird stets einem md. Cluster zugeschlagen, und die thüringische Handschrift B wird überhaupt keinem Cluster zugeordnet. Beispielhaft zeigt Abb. 6 das nd. Cluster; hier fehlt nur

---

[18] Der maximale Distanzwert von 6,5 kommt z.B. zwischen den Textzeugen SG (hchalem.) und f (östl. nnd.) vor. Einer der minimalen Pfade ist (in Klammern jeweils die bis dahin aufsummierten Distanzen): hchalem. (0) – schwäb. (0,5) – rhfrk. (2,5) – hess. (3,0) – wfäl. (5,0) – Rest-nnd. (6,0) – östl. nnd. (6,5).

[19] Die diplomatische Version bildet zu viele dokumentspezifische Eigenschaften ab. Bi- und Trigramme wiederum erfassen besser die charakteristischen Eigenschaften als Unigramme. Bei Wortart-N-Grammen ist allerdings die Datenmenge für Trigramm-Vergleiche zu klein.

das Fragment hb, das allerdings nur 55 Tokens umfasst und damit deutlich zu klein ist, um sinnvolle statistische Aussagen zu treffen.

– Innerhalb der Sprachräume der Stufe (ii) wird in allen Variationen die Gruppe der ripuarischen Textzeugen perfekt erfasst, s. Abb. 7. Der Druck N1500 zeigt dabei die meisten Eigenheiten.

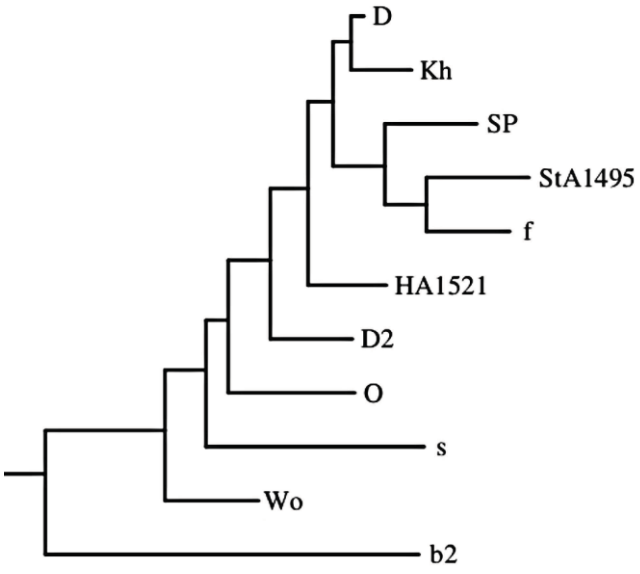

Abb. 6: Das nd. Cluster (Daten: simplifizierte Trigramme)

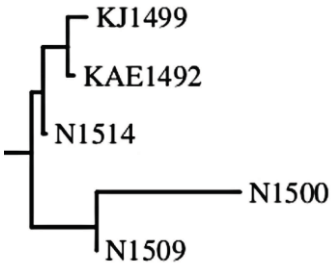

Abb. 7: Das rip. Cluster (Daten: simplifizierte Trigramme)

– Die große Gruppe der mittelbairischen Textzeugen unterteilt sich in mehrere kleinere Cluster. Dabei werden oft nordbairisch(-ostfränkische) und gelegentlich auch alemannische Textzeugen integriert.

– Zwischen den verschiedenen Parameter-Variationen sind die Unterschiede insgesamt eher klein und beschränken sich meist auf recht lokale Verschiebungen einzelner Dokumente. Beispielsweise bildet D2 bei den simplifizierten Bigrammen ein Cluster mit Kh (anders als in Abb. 5).

Nun zu den wortartbasierten Variationen: Insgesamt ist der Graph hier deutlich flacher, d.h. die Unterschiede zwischen den Clustern geringer. Die niederdeutschen Textzeugen bilden nach wie vor ein klares Cluster, gleiches gilt für die ripuarischen Dokumente. Die restlichen mitteldeutschen wie auch die oberdeutschen Dokumente tauchen jedoch verstreut in verschiedenen Clustern auf, d.h. die entstehenden Cluster sind sehr inhomogen, was die Sprachraumzugehörigkeit anbelangt. Unterscheidet man allerdings zusätzlich zwischen Vers- und Prosa-Fassungen, da dieser Faktor großen Einfluss auf die Syntax hat, so sieht man, dass sich alle mitteldeutschen Vers-Fassungen (D3 und b sowie alle ripuarischen Fassungen) tatsächlich innerhalb

eines Clusters befinden. Die Prosa-Fassungen hingegen sind verstreut auf heterogene Cluster.[20]

Abschließend seien noch die gravierendsten Abweichungen zwischen den buchstabenbasierten Clustern und den manuellen Zuordnungen genannt:

- Die mischklassifizierte Handschrift D4 (schwäbisch/rheinfränkisch) verhält sich generell eher wie die oberdeutschen Textzeugen.
- Die alemannische Handschrift St2 wird mehrfach mit mitteldeutschen Dokumenten gruppiert.

Schließlich scheinen einige Textzeugen recht idiosynkratische Eigenschaften aufzuweisen und tauchen daher regelmäßig am Ende dünn besiedelter Clusterzweige auf. Dies betrifft v.a. die Handschriften B (thüringisch), H (obersächsisch) und W (mittelbairisch).

3.3.2 Quantitativer Vergleich

Für den (automatischen) quantitativen Vergleich der Cluster nutze ich das Programm treedist von phylip mit der Methode ‚branch score distance'. Es berechnet Distanzen zwischen Topologien von Graphen (= unseren Clustern) und berücksichtigt dabei auch die Größen der Distanzen.[21] Die Variation, deren Graph die kleinste Distanz zu dem manuellbasierten Graph aufweist, sollte dann diejenige sein, die die Einschätzungen der Experten der historischen Linguistik am besten reproduzieren kann.

Zunächst kann man die Graphen der automatischen Variationen untereinander vergleichen. Hier zeigt sich, dass die beiden phonetischen Variationen (Bi- und Trigramme) sich am ähnlichsten sind (mit einem Distanzwert von 9,2).[22] Die nächstähnliche Paarung bilden die beiden simplifizierten Variationen (Bi- und Trigramme; Distanzwert 9,5). Vergleicht man die phonetischen mit den simplifizierten Variationen, so erhält man Werte zwischen 10,3 und 13,4. Am unähnlichsten zu allen anderen Graphen ist der Wortart-Graph (Bigramme; mit Distanzwerten zwischen 36,0 und 42,4).

---

20 Die oberdeutschen Textzeugen sind alle Prosa-Fassungen.

21 treedist bietet auch eine Methode (symmetric difference) an, die nur die Partitionierung der Cluster vergleicht und die Länge der Zweige ignoriert. Auf den ersten Blick scheint diese Methode besser geeignet für unsere Daten, da sie von den recht willkürlich gewählten Distanzen für die Sprachraumgrenzen abstrahiert. Da sich unsere Graphen jedoch oft durch kleine Verschiebungen innerhalb der Subcluster einer Sprachregion voneinander unterscheiden, die sich nicht negativ auswirken sollten, eignet sich dieses Vergleichsmaß weniger gut.

22 Hier und im Folgenden sind die Distanzwerte der besseren Lesbarkeit wegen jeweils mit 100 multipliziert.

Kommen wir nun zum Vergleich mit dem Graphen, der sich aus den manuellen Zuordnungen ergibt. Hier muss zunächst klargestellt werden, dass der manuelle Graph notwendigerweise deutlich größere Distanzen zu den automatischen aufweisen wird als die gerade genannten Distanzwerte, und zwar aus folgenden Gründen:

- Die Distanzwerte der manuellen Zuordnung sind eher willkürlich gewählt. Zwar liegen die beobachteten Maximalwerte der automatischen Variationen oft über 99% und damit sehr nahe am Maximalwert des manuellen Verfahrens, der bei 1 liegt. Bei den Minimalwerten ergeben sich aber beträchtliche Unterschiede. Z.B. treten beim manuellen Graphen Minimalwerte von 0 oder nahe 0 häufig auf, während beispielsweise bei den automatischen buchstabenbasierten Verfahren die beobachteten Minimalwerte über 65% lagen. Diese Differenz wird sich in entsprechend höheren Distanzwerten niederschlagen.
- Zudem sind die Kosten für Überquerungen von Sprachraumgrenzen (von 2, 1 und 0,5) recht zufällig gewählt und eher als relative denn als absolute Distanzen zu verstehen. Trotzdem nutze ich ein Graphen-Distanzmaß, das die Länge der Zweige berücksichtigt (zur Begründung s. Anm. 19).
- Die extrem kurzen Fragmente können als statistische Ausreißer das Bild verzerren. Die Fragmente stellen die Textzeugen, die die größten paarweisen Distanzen ergeben (s. Abschnitt 3.1).

Hier nun die Ergebnisse des Vergleichs der manuellen Zuordnung mit den automatischen Klassifikationen. Am ähnlichsten zum manuellen Graph ist die phonetische Bigramm-Variation (mit einem Distanzwert von 40,5), gefolgt von den simplifizierten Bigrammen (42,8), den phonetischen (43,7) und simplifizierten (45,9) Trigrammen. Einigermaßen abgeschlagen sind die Wortart-Bigramme (65,3).

Um eine Idee davon zu bekommen, wie diese Distanzwerte zu interpretieren sind, d.h. in welchem Maße die automatischen Graphen dem manuell-basierten (un-)ähnlich sind, habe ich fünf Zufallsgraphen generiert. Vergleicht man diese mit den anderen Graphen und untereinander, so liegen die Werte alle über 72,8, mit einem Durchschnittswert von 79,7 beim Vergleich der automatischen mit den Zufallsgraphen.[23] D.h. die buchstabenbasierten automatischen Verfahren erzeugen Cluster, die der manuellen Zuordnung recht ähnlich sind und deutlich besser abschneiden als die zufallsgenerierten Cluster. Bei der wortartbasierten Klassifikation ist der Unterschied zu den Zufallsgraphen hingegen weniger groß.

Um allerdings wirklich fundierte Aussagen machen zu können, müsste detailliert untersucht werden, worauf die Graph-Differenzen hauptsächlich beruhen. Neben den oben genannten Gründen können sich die Differenzen natürlich auch daraus ergeben, dass die automatischen Verfahren tatsächlich weniger gut geeignet sind. Die positiven Ergebnisse der qualitativen Auswertung sprechen aber eher gegen die-

[23] Die Ergebnisse im Einzelnen: Die fünf Random-Graphen verglichen mit den fünf buchstaben- und wortartbasierten Graphen ergeben eine durchschnittliche Distanz von 79,7 mit einer Standardabweichung von 6,2. Die Random-Graphen verglichen mit der manuellen Zuordnung ergeben 84,9±2,4, Random untereinander 105,0±2,7.

se Erklärung. Auf jeden Fall sollten in Nachfolgeuntersuchungen die Wahl der Distanzwerte für die manuelle Zuordnung optimiert und die Fragmente versuchshalber von den Vergleichen ausgeschlossen werden.

4. Zusammenfassung und Ausblick

In diesem Artikel spielten Variationen über ein Thema in zweifacher Hinsicht eine Rolle. Zum einen stellt der Anselm-Text selbst mit seinen z.T. sehr verschiedenen Textzeugen das Thema mit Variationen dar. Zum anderen habe ich die Textzeugen auf verschiedene Arten transformiert und variiert und mich dabei immer weiter vom Original entfernt. Den weitesten Weg gehen die Variationen in Wortart-Form.

Das Anselm-Korpus stellt, dank seiner breiten Überlieferung aus allen wesentlichen Sprachregionen des deutschen Sprachraums, eine wunderbare Quelle für sprachvergleichende Arbeiten dar. In diesem Artikel wollte ich einige der computerlinguistischen Möglichkeiten zeigen, um Sprachvergleiche automatisch durchzuführen. Diese Methoden können v.a. dazu dienen, auf potenziell interessante Zusammenhänge aufmerksam zu machen, die dann in einem zweiten Schritt manuell untersucht werden können. In diesem Artikel habe ich versucht, bereits bekannte Ergebnisse, die am Lehrstuhl Wegera entstanden sind, zu replizieren.

Der vorgestellte Ansatz ist in vielerlei Hinsicht noch verbesserungsbedürftig. Beispielsweise fehlt eine Analyse der Fehlerraten der automatischen Tools, so dass unklar ist, ob bestimmte Fehlklassifizierungen auf fehlerhaften Input zurückführbar sind oder ob die Methode an sich nur bedingt für die Aufgabe geeignet ist. Außerdem könnte (und sollte) man die Analyse-Tools für historische Sprachdaten optimieren. Schließlich habe ich bei den automatischen Vergleichen verschiedene wichtige Faktoren ignoriert, wie z.B. die Textgattung (Vers vs. Prosa), die v.a. relevant sein sollte für die wortartbasierten Vergleiche, oder die zeitliche Einordnung der Textzeugen. Letzteres habe ich wegen der geringen Datenmenge nicht berücksichtigt.

Nicht zuletzt legt die quantitative Auswertung nahe, kurze Fragmente auszuschließen und die Distanzmaße der manuellen Zuordnung zu überdenken.

Trotz all dieser Einschränkungen lässt sich m.E. jedoch festhalten, dass die Ergebnisse insgesamt positiv ausfallen und die computerlinguistischen Methoden vielversprechend für die historisch-linguistische Forschung zu sein scheinen. Als Gewinner des Variationswettbewerbs ergeben sich aus der quantitativen Auswertung die beiden bigrammbasierten Variationen (simplifiziert und phonetisch).