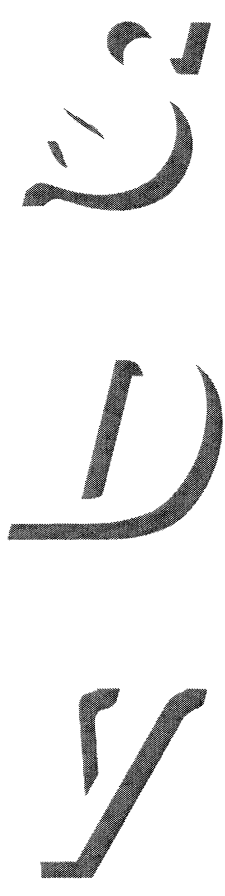
Sprache und Datenverarbeitung

International Journal for Language Data Processing

31. Jahrgang 2007 Heft 1-2



Begründet durch
Winfried Lenders und Harald Zimmermann

Herausgegeben durch das
Institut für Kommunikationswissenschaften
der Universität Bonn
Abteilung Sprache und Kommunikation
Poppelsdorfer Allee 47
53115 Bonn

von:
Hermann Cölfen, Essen
Annelie Rothkegel, Chemnitz
Ulrich Schmitz, Essen
Bernhard Schröder, Essen

Schriftleitung:
Ulrich Schmitz
Universität Duisburg-Essen
Fachbereich Geisteswissenschaften
Universitätsstraße 12
45117 Essen
E-Mail: ulrich.schmitz@uni-due.de

Layout:
Sabine Walther, Duisburg

Titelillustration:
Michael Hüter, Bochum

Sprache und Datenverarbeitung im Internet: http://www.linse.uni-due.de

Thomas Klein

Zur Dimensionierung historischer Textkorpora

Abstract English

Annotated historical linguistic corpora need to be restricted in order to create a corpus large enough for the intended analysis on the one hand but not too large to hinder the realisation of the project on the other hand. The following paper will show that data from a smaller comparable corpus is required in order to optimize the scope of the new corpus.

Abstract Deutsch

Der Umfang annotierter historischer Sprachkorpora muss so beschränkt werden, dass das Korpus einerseits für die intendierten Untersuchungen groß genug ist und andererseits nur so groß, dass Erstellung und Annotation des Korpus realisierbar sind. Der folgende Aufsatzes will zeigen, dass man für die Festlegung des optimalen Korpusumfangs schon über die Daten aus einem (kleineren) Vergleichskorpus verfügen muss.

Die Notwendigkeit, ein Textkorpus zu bilden, und die Frage, wie es zu strukturieren sei, ergeben sich bei historischen Sprachstufen erst dort, wo die erhaltene Gesamtüberlieferung so reichlich ist, dass sie das Maß des aktuell Bearbeitbaren überschreitet. Bei nur ganz spärlich überlieferten Sprachen oder Sprachstufen[1] wie etwa dem Krimgotischen oder der Sprache der älteren Runeninschriften, doch auch bei Sprachen mit begrenztem Überlieferungsumfang wie etwa dem Bibelgotischen sollte das Textkorpus, das die Untersuchungsbasis für diese Sprache bildet, mit der Gesamtüberlieferung zusammenfallen. Für die älteren Sprachstufen des Deutschen ist dies für den Zeitraum bis etwa 1200 gleichfalls grundsätzlich möglich und erstrebenswert.[2] Ab dem 13. Jahrhundert schwillt die Gesamtüberlieferung des Mittelhochdeutschen und mehr noch die des sich anschließenden Fühneuhochdeutschen dann aber derart an, dass die Bildung eines Auswahlkorpus aus Gründen der Machbarkeit und Finanzierbarkeit unerlässlich ist.

Doch es gibt auch gewichtige Sachgründe für die Beschränkung auf eine Textauswahl, insbesondere dann, wenn es um ein strukturiertes Textkorpus geht, wie es für viele korpuslinguistische Untersuchungen erforderlich, zumindest aber wünschenswert ist. Ein

1 Vgl. dazu Untermann (1989) und die weiteren Aufsätze in Beck (1989).

2 Dies ist denn auch das Ziel der in engem Verbund stehenden Projekte „Annotiertes Corpus Altdeutsch (ACA)“ und „Annotiertes Corpus Mittelhochdeutsch (ACM)“.

solches strukturiertes Auswahlkorpus bildet die Textbasis der entstehenden neuen Mhd. Grammatik. Im Folgenden soll dieses Korpus als Ausgangspunkt für die Betrachtung von Fragen dienen, die sich bei der Festlegung des Umfangs eines sprachhistorischen Textkorpus stellen.

Das Mhd. Grammatik-Korpus, das auf dem 1993-1995 unter Leitung von Klaus-Peter Wegera erstellten „Bochumer Mittelhochdeutsch-Korpus" basiert, setzt sich aus etwas über 100 Texten oder Textausschnitten zusammen (vgl. Wegera 2000). Sie verteilen sich so gleichmäßig über ein zeitlich-sprachräumliches Raster, wie es die vor allem in frühmhd. Zeit leider nur sehr spärliche Überlieferung gestattet.

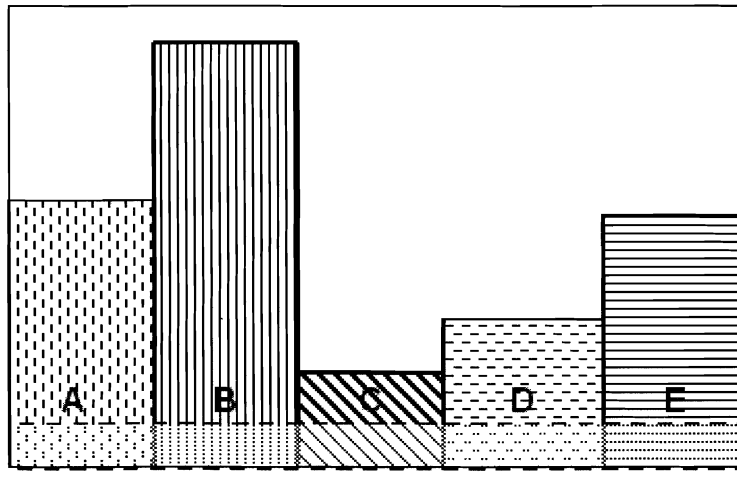


Abb. 1: Einebnung distributioneller Unterschiede innerhalb der Gesamtüberlieferung im strukturierten Korpus

Ein konstruktionelles Grundprinzip des Bochumer Mittelhochdeutsch-Korpus – und in seinem Gefolge auch des Mhd. Grammatik-Korpus – ist es somit, dass die verschiedenen Zeitabschnitte, Sprachräume und Textsorten im Korpus soweit wie möglich nach Text und Textumfang gleichgewichtig vertreten sind, um eine adäquate Grundlage für vergleichende diaphasische, diatopische und textsortenbezogene Untersuchungen zu schaffen. Das bedeutet, dass distributionelle Verzerrungen[3] in der Überlieferungsgesamtheit im Korpus korrigiert, nämlich zwecks einer Gleichgewichtung eingeebnet werden (s. Abb. 1). Damit ist freilich nur entschieden, dass die Segmente der Korpusstruktur mit der gleichen Anzahl von Texten gefüllt sein sollten, nicht aber, wie groß diese Zahl zu sein habe.

Ähnliche Fragen stellen sich hinsichtlich des Textumfangs, der natürlich extrem unterschiedlich sein kann. Beispielsweise umfasst der ‚Erfurter Judeneid' lediglich 120

3 Solche Verzerrungen könnten beispielsweise dadurch entstanden sein, dass die mhd. Überlieferung eines Sprachraums durch Kriege und andere ungünstige Umstände in nachmhd. Zeit weit stärker reduziert worden ist als die eines anderen Sprachraums.

Wortformen (= Tokens), das ‚Passional' dagegen 550.000 Wortformen. Während 120 Wortformen viel zu wenig sind, um abgesicherte Aussagen zu erlauben, sind 550.000 Wortformen pro Einzeltext zum einen unter realistischen Rahmenbedingungen nicht bearbeitbar, zum andern überschreiten sie für die meisten grammatischen Fragestellungen das Maß des Notwendigen um ein Vielfaches. Auch hier wird man sich daher schon aus auswertungspraktischen Gründen für einen möglichst gleich großen Umfang der Texte bzw. Textausschnitte entscheiden. Wie groß muss dieser Umfang aber sein, damit das Korpus seinen Zweck erfüllen kann? Zur Beantwortung dieser Frage müsste man offenbar abschätzen können, wieviel Belege der linguistischen Objekte, deren Merkmale man untersuchen will, bei einem bestimmten Textumfang zu erwarten sind. Kennt man nämlich die Zahl b der Belege eines Objekts bei einem Textumfang von n Wortformen, dann lässt sich schätzungsweise auch der Textumfang n' bestimmen, bei dem sich eine hinreichende[4] Belegzahl b' ergibt ($n' = b' \cdot n/b$).[5] Unterschreitet man dieses Maß, so sind Erstellung und Annotierung des Korpus zwar leichter realisierbar, doch wäre das Korpus für die Untersuchungszwecke, für die es gedacht ist, allenfalls eingeschränkt tauglich. Überschreitet man das Maß dagegen in Richtung auf ein sehr großes und entsprechend sehr gut geeignetes Korpus, so werden Korpuserstellung und -annotierung zunehmend schwerer realisierbar und finanzierbar (s. Abb. 2).

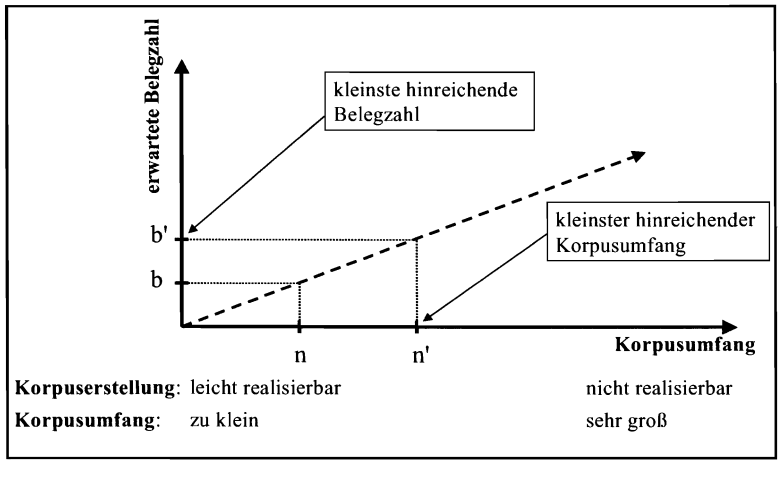


Abb. 2: Korpusdimensionierung nach der zu erwartenden Belegzahl

Der Schluss, der aus dieser Überlegung zu ziehen ist, lautet offenbar: Um ein Korpus angemessen zu dimensionieren, muss man bereits über ein vergleichbares Korpus verfügen, an dem sich bei gegebenem Korpusumfang n die Belegzahl b berechnen lässt.

4 Welche Belegzahl als hinreichend anzusehen ist, welcher Stichprobenumfang also optimal ist, hängt dann von dem benutzten statistischen Testverfahren ab und ist hier nicht weiter zu betrachten.

5 Vorausgesetzt wird dabei ein lineares Wachstum der Belegzahl bei wachsender Textlänge.

Vergleichbar wäre ein Korpus, das derselben Sprachstufe angehört und möglichst ähnlich strukturiert ist.

Das in Abb. 2 dargestellte Verfahren bedeutet allerdings eine starke Vereinfachung, da es die Beleglage nur eines linguistischen Untersuchungsobjekts für die Korpusdimensionierung berücksichtigt, während Textkorpora in aller Regel für die Untersuchung vieler Objekte und ihrer Merkmale mit sehr unterschiedlichen Vorkommenshäufigkeiten genutzt werden sollen. Das erschwert die Ermittlung des optimalen Korpusumfangs ganz erheblich.

Im Projekt der Grammatik des Frühneuhochdeutschen, das in den 1970er und 1980er Jahren in Bonn und Augsburg die Flexionsmorphologie des Frühneuhochdeutschen untersucht hat, wurde die Textlänge von 30 Normalseiten à 400 Wortformen, also 12.000 Wortformen, als für diesen Untersuchungszweck ausreichend festgelegt. Dieser Wert, der zwar nicht systematisch ermittelt wurde, sich aber in der Praxis bewährt hat, ist auch für das Bochumer Mittelhochdeutsch-Korpus und damit für das Mhd. Grammatik-Korpus übernommen worden. Da das Grammatik-Korpus in der Bonner Arbeitsstelle des Projekts Mhd. Grammatik komplett lemmatisiert und grammatisch annotiert worden ist, lässt sich das 12000-Wortformen-Maß nun sehr genau auch für einzelne Einheiten und ihre Merkmale überprüfen. Erwartungsgemäß sind die verschiedenen Flexionsformen sehr unterschiedlich häufig belegt. Für die finiten Verbalformen (ohne Berücksichtigung der Modusunterschiede) ergeben sich im Grammatik-Korpus die in Abb. 3 dargestellten Verhältnisse.

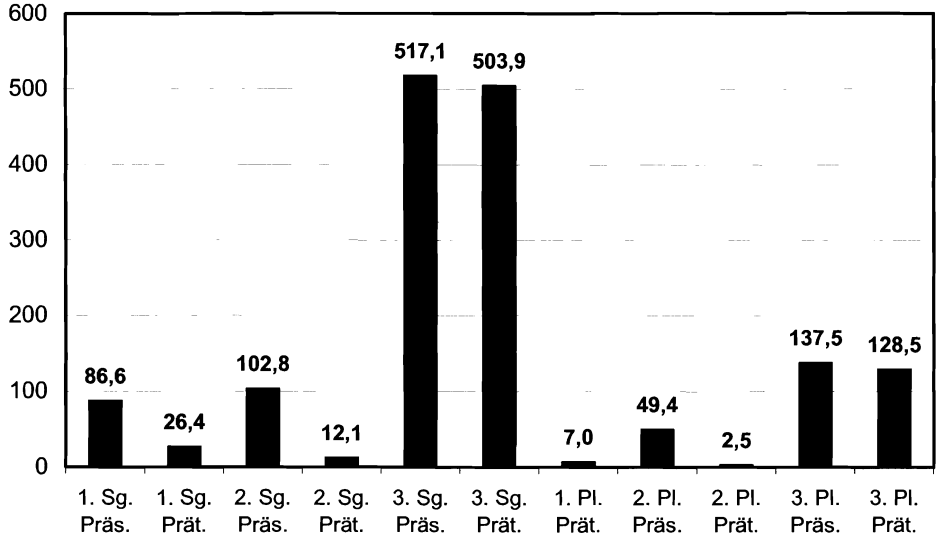


Abb. 3: Häufigkeit von Verbformen pro 12000 Wortformen im Korpus der Mhd. Grammatik

Der Textumfang von 12000 Wortformen genügt demnach für die meisten finiten Verbformen. Für einige – insbesondere die 3. Sg. Präs. und Prät. – wäre schon eine weit geringere Textlänge ausreichend. Demgegenüber kommen andere in 12000 Wortformen im Durchschnitt nur so selten vor (vgl. Abb. 4), dass zumindest bezogen auf den Sprachstand des Einzeltextes keine verlässlichen Aussagen über diese Flexionsformen möglich sind.

Ein gutes Beispiel für das Problem ist die 2. Pl. Prät.: Um beurteilen zu können, inwieweit in dem betreffenden Text die Flexionsendung *-(e)nt* der 2. Pl. Prät. gilt, reichen die wenigen Belege nicht aus. Sollte daraus der Schluss gezogen werden, dass die Wortformzahl pro Text wegen dieses einen Sprachmerkmals entsprechend erhöht werden müsste? Das wäre hochproblematisch. Wollte man die Belegzahl pro Text global von 2,5 auf etwa 10 Belege steigern, so würde sich der Gesamtaufwand für die Korpuserstellung und -annotation vervierfachen. Das wäre im Rahmen des Projekt Mhd. Grammatik nicht zu leisten gewesen (zumindest aber wäre die erheblich vermehrte Personalausstattung und Laufzeit von der DFG schwerlich finanziert worden).

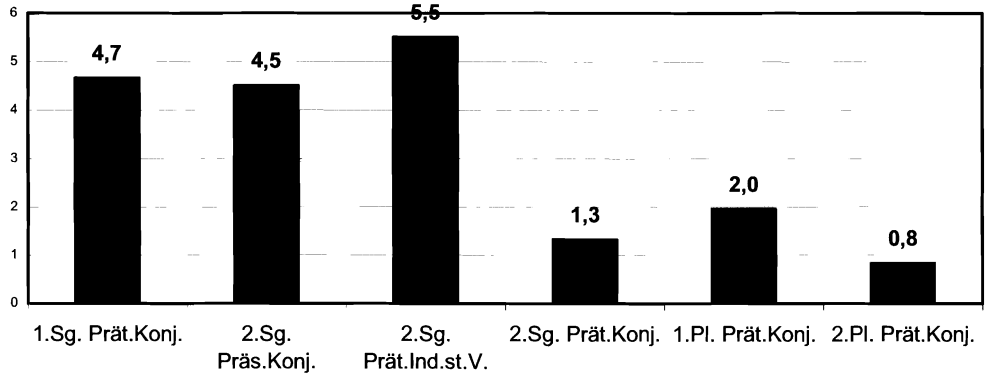


Abb. 4: Häufigkeit seltener Verbformen pro 12.000 Wortformen im Korpus der Mhd. Grammatik

Eine Alternative zur globalen Korpuserweiterung wäre jedoch eine regional beschränkte. Die 2. Pl. auf *-(e)nt* ist bekanntlich eine vorzugsweise alemannische Erscheinung, die aber auch ins Rheinfränkische und Westbairische hineinreicht.[6] Man könnte daher gezielt den Korpusausschnitt des (weiteren) Südwestens durch Hinzunahme zusätzlicher Texte und durch die Vergrößerung der Ausschnitte aus bereits berücksichtigten Großtexten erweitern. Das verspräche bei erheblich geringerem Aufwand im Wesentlichen denselben Erkenntnisgewinn wie die globale Korpuserweiterung – allerdings mit dem Risiko, einschlägige Belege außerhalb des gewählten Raumausschnitts zu übersehen.[7] Auch

6 Vgl. Paul 2007, § M 70 Anm. 8. Diese sprachgeographische Verteilung bestätigt sich auch in den einschlägigen Belegen des Mhd. Grammatik-Korpus.

7 So etwa die Vorkommen der 2. Pl. auf *-ent* bei den hochdeutsch schreibenden Niederdeutschen, die freilich einen Sonderfall bilden.

bei der Belegermittlung innerhalb des Zusatzkorpus sollte anders verfahren werden als im Gesamtkorpus: Statt einer zeitraubenden kompletten Annotation des Zusatzkorpus könnten gezielt nur die Belege der 2. Pl. über die Vorkommen der Form *ir* (nebst Varianten wie *jr*, *Ir*, *Jr*, *ier* etc.) ermittelt werden.[8]

Das Beispiel verdeutlicht, dass bei einer merkmalsbezogenen Korpusdimensionierung nicht die sehr selten belegten Merkmalsträger den Ausschlag geben sollten. Für sie können vielmehr spezielle Wege gesucht werden, die zielführend und doch unaufwändig gehbar sind.

Dies gilt nicht nur für einzelne Einheiten und ihre Merkmale, sondern ebenso für ganze Sprachebenen. Bei gleichem Textumfang kommen die Einheiten der verschiedenen Sprachebenen bekanntlich sehr unterschiedlich häufig vor: Graphien wie <t> sehr viel häufiger als Morphe wie *-et* und diese wiederum viel häufiger als syntaktische Einheiten wie Relativsätze. Schon aus dieser Selbstverständlichkeit ergibt sich, dass der optimale Korpusumfang sprachebenenbezogen unterschiedlich sein wird (s. Abb. 5).

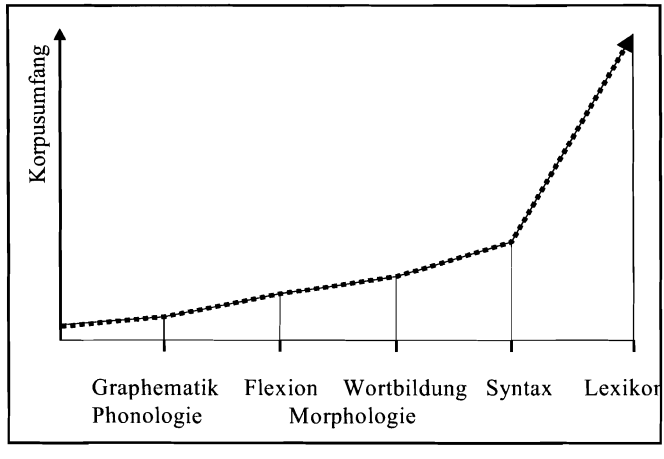


Abb. 5: Sprachebenenbezogene Korpusdimensionierung

Wie der Korpusumfang für eine bestimmte Sprachebene aber konkret auszufallen habe, lässt sich wiederum nur dann begründet entscheiden, wenn Daten aus einem Vergleichskorpus vorliegen. Für den graphematisch-phonologischen Bereich reichen im Mhd. schon ca. 3000 Wortformen im Schnitt aus, damit sich der Lautstand eines Textes hinreichend scharf abzeichnet. Insbesondere die meisten für die sprachgeographische Einordnung relevanten Laute oder Lautpositionen kommen bei diesem Textumfang schon recht häufig vor.[9] Erwartungsgemäß nur selten bis sehr selten belegt sind dagegen

8 Man würde sich dabei zunutze machen, dass Formen der 2. Pl. in aller Regel nur in der Umgebung des Subjektpronomens *ir* vorkommen, dessen Belege freilich zunächst von denen des Gen./Dat. Sg. Fem., Gen. Pl. oder Poss.-Pronomens *ir* zu trennen wären.

9 Z. B. sind an Konsonanten pro 3000 Wortformen belegt: im Anlaut: 67,4 *k*, 109,7 *b*, 67 *t*, zwischenvokalisch 76,9 *t*, 67,7 *b*, im Auslaut nach Vokal 33,6 *g*; noch knapp ausreichend sind belegt: nach Konsonant 13,5 *k*, 19,3 *t* im Inlaut nach *l*, nach *r* 13,1.

die Nachfolger von vorahd. **p* im Anlaut und nach Konsonant und die Nachfolger der vorahd. Geminaten **pp*, **tt*, **kk*, **bb*, **dd*; nur selten auch mhd. *v* (< vorahd. **f*) im In- und Auslaut:

Laut(position)	**Belege**	
	pro 3000 Wortformen	**pro 12000 Wortformen**
mhd. *pf* im Anlaut	7,53	30,11
mhd. *pf* nach Nasal	0,43	1,73
mhd. *f* < vorahd. **p* nach Liquid	4,31	17,25
mhd. *pf* < vorahd. **pp*	1,74	6,96
mhd. *ck* < vorahd. **kk*	6,18	24,72
mhd. *pp* < vorahd. **bb*	1,05	4,19
mhd. *tt* < vorahd. **dd*	8,30	33,20
mhd. *tz* < vorahd. **tt*	6,60	26,39
mhd. *v* zwischen Vokalen	6,89	27,57

In zwei Drittel dieser Fälle würde sich die Belegzahl bei einem Textumfang von 12.000 Wortformen hinreichend erhöhen, aber gerade bei den sprachgeographisch wichtigen Kriterien mhd. *pf* nach Nasal und mhd. *pf* < vorahd. **pp* würde auch das nicht viel fruchten. Wieder stellt sich die Frage, ob man aus Rücksicht auf wenige selten belegte Einheiten den Korpusumfang und damit den Bearbeitungsaufwand vervielfachen sollte. Auch hier ist es m. E. besser, sich für den leichter zu bewältigenden geringeren Textumfang zu entscheiden und für die dann zu selten belegten Erscheinungen arbeitsökonomische spezielle Lösungen zu suchen.[10]

Die vorstehenden Bemerkungen sollten zeigen, dass bereits die Festlegung des Umfangs eines historischen Textkorpus ein erhebliches Maß an Vorwissen verlangt und dass dieses Vorwissen zu einem guten Teil nur über die Auswertung eines schon bestehenden Vergleichskorpus zu gewinnen ist. Korpora historischer Sprachstufen sollten auch von daher nicht als starre, fest abgeschlossene Größen, sondern als dynamische Gebilde betrachtet werden, die sukzessive zu verbessern und auszubauen eine fortdauernde Aufgabe bleibt.[11]

10 Gerade bei mhd. *pf* liegen solche Lösungen nahe, so vor allem die Vergrößerung des Textumfangs nur bei mitteldeutschen Texten und – da sich etwa *pf* nach Nasal auf ganz wenige Lexeme und ihre Ableitungen beschränkt – lexemspezifische Volltextrecherchen zur Belegermittlung.

11 Dafür ist auch die Entwicklung das Mhd. Grammatik-Korpus ein gutes Beispiel. Es hat sich vor allem in den Jahren 1996-98 während der Arbeit an dem DFG-Projekt „Korpus einer mhd. Grammatik" (Th. Klein – H.-J. Solms – K.-P. Wegera) in Textbestand und Struktur stark gewandelt und auch seither noch in einer ganzen Reihe von Punkten verändert. Vergleicht man es mit seiner Basis, dem Bochumer Mittelhochdeutschkorpus von 1995, so ist jeder zweite Text von diesen Veränderungen betroffen, sei er nun entfernt oder ersetzt worden oder neu hinzugekommen oder innerhalb der Korpusstruktur an einen andern Platz gelangt.

Literatur

Beck, Heinrich (Hrsg.) (1989): Germanische Rest- und Trümmersprachen. Berlin, New York (Ergänzungsbände zum Reallexikon der Germanischen Altertumskunde, Bd. 3).

Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre hrsg. von Hugo Moser, Hugo Stopp u. Werner Besch. Bd. III: Flexion der Substantive von Klaus-Peter Wegera. Heidelberg 1987; IV: Flexion der starken und schwachen Verben von Ulf Dammers, Walter Hoffmann, Hans-Joachim Solms. Heidelberg 1988; VI: Flexion der Adjektive von Hans-Joachim Solms, Klaus-Peter Wegera. Heidelberg 1991; VII: Flexion der Pronomina und Numeralia von Maria Walch, Susanne Häckel. Heidelberg 1988.

Paul, Hermann (2007): Mittelhocheutsche Grammatik. 25. Aufl. neu bearb. von Thomas Klein, Hans-Joachim Solms u. Klaus-Peter Wegera. Mit einer Syntax von Ingeborg Schröbler, neubearb. u. erweitert von Heinz-Peter Prell. Tübingen.

Untermann, Jürgen (1989): Zu den Begriffen ‚Restsprache' und ‚Trümmersprache'. In: Beck (1989), 15-19.

Wegera, Klaus-Peter (2000): Grundlagenprobleme einer neuen mittelhochdeutschen Grammatik. In: Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 2. Auflage. 2. Teilband. Hg. von Werner Besch, Anne Betten, Oskar Reichmann, Stefan Sonderegger (Handbücher zur Sprach- u. Kommunikationswissenschaft, Bd. 2.3). Berlin / New York, 1304–1320.