Originalveröffentlichung in: Jahrbuch für germanistische Sprachgeschichte Jg. 10 (2019), Nr. 1, S. 1-28

Kristin Kopf

# Von Korpus zu Korpus

## Herausforderungen und Chancen diachron korpusübergreifenden Arbeitens

**Abstract:** Das Kombinieren von Daten aus verschiedenen diachronen Korpora bringt besondere methodische Herausforderungen mit sich, die in den vorliegenden Untersuchungen beleuchtet werden. Dazu gehört der Abgleich von Metadaten und ihrer Kategorisierungen, das Verhalten bekannter Phänomene über sich zeitlich überschneidende Korpora hinweg und die Formulierung vergleichbarer Suchabfragen. Anhand von sechs Fallstudien zu graphematischen, lexikalischen, morphologischen und syntaktischen Phänomenen in Korpora des (Früh-) Neuhochdeutschen werden Möglichkeiten und Probleme des diachron korpusübergreifenden Arbeitens herausgearbeitet.

**Keywords:** Korpus, Sprachwandel, Frühneuhochdeutsch, Neuhochdeutsch, Deutsches Textarchiv, Bonner Frühneuhochdeutschkorpus, GerManC, Graphematik, Virgel, Diphthong, Deklinationswandel, Indefinitartikel

## 1 Einleitung

Daten aus verschiedenen diachronen Korpora werden schon lange kombiniert und gemeinsam ausgewertet. Dabei kommt es jedoch bisher kaum vor, dass eine Untersuchung über mehrere Korpora hinweg konzipiert wird. Das dürfte am doppelten Problem der Vergleichbarkeit liegen: Zum einen sind die Korpora sehr heterogen zusammengesetzt. Zum anderen verfügen sie über unterschiedliche Abfragemöglichkeiten und -werkzeuge, die ihre eigenen Herausforderungen besitzen. Insbesondere die Konzeption vergleichbarer Suchabfragen erfordert weitreichende Entscheidungen: Grammatische Kategorien wie z. B. Kasus lassen sich in kleineren Korpora gezielt abfragen, während etwa im Deutschen Textarchiv (DTA) Umwege über den Kontext oder eine Beschränkung auf Einzellexeme nötig sind. Suchabfragen müssen aber so konzipiert werden, dass in allen Korpora

**Kristin Kopf,** Westfälische Wilhelms-Universität Münster, Germanistisches Institut, Schlossplatz 34, 48146 Münster und Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, E-Mail: kristin.kopf@uni-muenster.de

https://doi.org/10.1515/jbgsg-2019-0004

dasselbe gefunden wird. Für die Suchabfragen muss daher ein Kompromiss zwischen der Nutzung korpusspezifischer Informationen und möglichst großer Gemeinsamkeit gefunden werden.

Die vorliegenden Untersuchungen stellen sich den angesprochenen Herausforderungen mit dem Ziel, Erkenntnisse über die Vergleichbarkeit diachroner Korpora zu gewinnen. Zum bisherigen, i.d.R. impliziten Umgang mit Daten verschiedener Korpora finden sich einige knappe Überlegungen in Kap. 2. Kap. 3 liefert eine intensive Analyse der potenziellen Datengrundlage und geht dabei über den Rahmen der folgenden Fallstudien hinaus. Es kann damit auch als Beschreibung der aktuell verfügbaren größeren diachronen Korpora des Deutschen alleine stehen. Die durchgeführten Fallstudien verteilen sich schließlich auf zwei Kapitel: In Kap. 4 werden mehrere bereits gut erforschte Phänomene (Virgelgebrauch, Schreibung des Diphthongs /a͡ɪ/, Gebrauchsfrequenz zweier Lexeme) in drei Korpora des (Früh-)Neuhochdeutschen getestet (Bonner Frühneuhochdeutschkorpus, GerManC, DTA). Ziel ist dabei nicht, neue Erkenntnisse zu den Phänomenen zu erhalten. Es wird im Gegenteil davon ausgegangen, dass sie sich genau so verhalten wie bisher in der Literatur beschrieben. Deshalb eignen sie sich dazu, das Verhalten der Korpora selbst zu überprüfen, sie quasi zu „eichen“: Folgen die Daten eines Korpus nicht dem bisher beschriebenen Muster (hier immer diachrone Zu- oder Abnahme) oder unterscheiden sich die Werte sich zeitlich überschneidender oder berührender Korpora stark, so wird geprüft, ob und ggf. wo es Probleme in der Zusammensetzung eines der Korpora gibt.[1] In Kap. 5 werden zwei Phänomene vorgestellt, deren Entwicklungsrichtung zwar ebenfalls bereits bekannt ist (Stammerweiterung schwacher Maskulina, Ausbreitung des Indefinitartikels), für die jedoch auf korpusspezifische Annotationen zurückgegriffen wird. Hieran lässt sich das Problem korpusübergreifend vergleichbarer Suchanfragen verdeutlichen.

## 2 Umgang mit diachronen Daten

Korpusübergreifendes Arbeiten wird für das Deutsche methodisch selten diskutiert und reflektiert. In bisherigen Studien finden sich drei grundsätzliche Ansätze für den Umgang mit diachronen Korpusdaten aus zeitlich angrenzenden oder sich überschneidenden Korpora: Eine Art Metaanalyse publizierter For-

1 Natürlich müssen später prinzipiell auch Stellen genauer analysiert werden, an denen Daten übereinstimmen; das kann immer auch einem Zufall geschuldet sein. Der vorliegende Aufsatz versteht sich nur als erster Schritt und konzentriert sich zunächst auf die offensichtlicheren Aspekte.

schungsergebnisse, vollständig eigene Erhebungen oder die Kombination von beidem, also die Ergänzung von Korpusbefunden aus der Literatur durch eigene Daten (z.B. bei Kempf 2016: 128). Versuche, Metaanalysen durchzuführen, kranken i.d.R. daran, dass die zugrundeliegenden Daten selten publiziert sind und entsprechend auch keine einheitliche Umcodierung erfolgen kann. In vielen Fällen, insbesondere bei älteren Studien, fehlen auch sonstige Angaben zum Untersuchungsrahmen (z.B. was aussortiert wurde, nach welchen Kriterien Kategorien gebildet wurden). So liefert z.B. eine Analyse publizierter Daten zum Stellungswandel des attributiven Genitivs vom Ahd. bis ins 18. Jh. bei Kopf (2018: 88) keine Erkenntnisse zur Diachronie des Phänomens über die banale Feststellung hinaus, dass sich irgendwann die Nachstellung durchsetzt (zu diesem Problem vgl. auch Pickl in diesem Band).

Erkenntnisreicher ist die Kombination eigener Erhebungen aus verschiedenen diachronen Korpora. Dabei können die Daten als separate Datensätze behandelt (z.B. Hartmann 2016: 178) oder als gemeinsamer Datensatz angesehen werden. Im zweiten Fall ist sowohl eine einheitliche Darstellung und Visualisierung (z.B. Christiansen 2016, der de facto ein neues Korpus schafft) als auch eine optische Trennung der Daten innerhalb einer Grafik (Kopf 2017: 185, 198 mit drei Korpora) möglich. Die getrennte Darstellung und Analyse ist methodisch zwar zunächst sauber, weil der Aspekt der Korpusvergleichbarkeit weitgehend ausgeklammert werden kann, allerdings werden Leserinnen und Leser doch zumeist versuchen, eine Entwicklung in die Daten zu lesen. Eine explizite Thematisierung dieses Aspekts ist also in jedem Fall notwendig.

Die Ziele bei der Kombination verschiedener Daten können unterschiedlicher Natur sein. So lassen sich durch den Vergleich sich zeitlich überlagernder Zeitschnitte (möglicherweise ungewöhnlich scheinende) Daten validieren (so z.B. bei Kempf 2016: 128, Kopf 2017), Artefakte in einem Korpus oder Zeitschnitt können identifiziert werden. Oft wird die Kombination außerdem nötig, weil sich z.B. herausstellt, dass vor oder nach dem Untersuchungszeitraum Relevantes passiert sein muss, das gewählte Korpus also eigentlich nicht die richtige Zeitspanne umfasst (z.B. Kopf 2017, einger.). Und schließlich soll ein Phänomen oft in seiner kompletten Diachronie beleuchtet werden, was ein einzelnes Korpus nicht leisten kann.

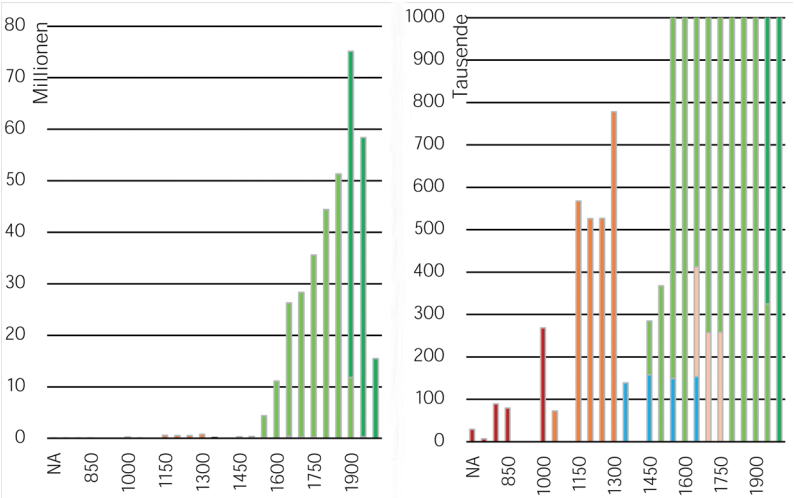
In all diesen Fällen steht jedoch zumeist das konkrete Phänomen im Zentrum, während die Methoden, die Zulässigkeit und mögliche Probleme der Datenkombination nur selten thematisiert werden. Hierzu will der vorliegende Aufsatz einen Beitrag leisten.

# 3 Diachrone (Referenz-)Korpora des Deutschen

Im Folgenden wird ein knapper Überblick über die derzeit verfügbaren, größeren diachronen Korpora des Deutschen gegeben. Neben Informationen zu ihrem relativen und absoluten Umfang und ihrer linguistischen Aufbereitung steht vor allem der Umgang mit Metadaten (Region und Textsorte) im Zentrum: Die Korpora weisen hier keine deckungsgleichen Kategorien auf, sondern müssen, falls erforderlich, nachträglich selbst in Übereinstimmung gebracht werden.

## 3.1 Überblick und Umfang

Den kleinen, sorgfältig und fein annotierten Korpora der älteren Sprachstufen stehen die großen Textsammlungen der DWDS-Kernkorpora (20. und 21. Jh.) und des DTA (15. bis 20. Jh.) gegenüber.




| Korpus | Tokens | Text(ausschnitt)e |
|---|---|---|
| Referenzkorpus Altdeutsch | [2]472.515 | (124) |
| Referenzkorpus Mittelhochdeutsch (ReM) | 2.470.517 | 394 |
| Bonner Frühneuhochdeutschkorpus (Bonner) | 599.695 | 40 |
| GerManC | [3]774.380 | 338 |
| Deutsches Textarchiv (DTA) | 213.028.633 | 4.393 |
| DWDS-Kernkorpus 20, 21 (DWDS) | 136.859.901 | 91.300 |

**Abbildung 1 und 2:** Größe diachroner Korpora des Deutschen in Textwörtern (Stand September 2018). Rechts Ausschnitt mit kleinerer Skala.

Wie in Abbildung 1 deutlich wird, weist das DTA allein in der zweiten Hälfte des 16. Jhs. schon einen größeren Umfang auf als das gesamte Bonner Frühneuhochdeutschkorpus. In Abbildung 2 wurde die Skala bei 1 Mio. Textwörtern abgeschnitten, um auch die Größen der älteren Korpora sichtbar zu machen.

Von besonderem Interesse sind für die vorliegende Untersuchung Korpora, die zeitliche Überschneidungen miteinander aufweisen, d.h. Bonner Korpus und GerManC als kleine Korpora und DTA als große Textsammlung.[4] In einer der im Folgenden vorgestellten Fallstudien werden auch Daten aus ReM und Altdeutschkorpus berücksichtigt. Ein detaillierter, systematischer Vergleich von Daten aus Altdeutschkorpus und ReM bleibt vorläufig aber noch ein Desiderat.

## 3.2 Zusammensetzung

Neben der Größe und dem abgedeckten Zeitraum ist die Zusammensetzung eines Korpus nach Zeitschnitten, Dialekträumen/Regionen und Textsorten für korpusübergreifende Untersuchungen von besonderer Relevanz.

Alle Korpora sind intern zeitlich gegliedert (Tabelle 1), allerdings nicht identisch. Um Vergleichbarkeit herzustellen, werden die Daten in den vorliegenden Studien in Zeitschnitte von 50 Jahren zusammengefasst.[5] Dialektraum und Textsorte bedürfen einer nachträglichen Kategorisierung: Für DTA und DWDS fehlen Angaben zum Raum. Sie können mit etwas Aufwand aus dem Verlagsort extrahiert werden.[6] Den kleinsten gemeinsamen Nenner stellt die Aufteilung von GerManC dar, für die Analyse in Abbildung 3–6 werden West- und Ostoberdeutsch zusätzlich zu Oberdeutsch zusammengefasst.

---

**2** Nur Althochdeutsch; lateinische und altsächsische Texte/Textpassagen wurden nicht mitgezählt. (Das Korpus ist auf Tokenebene nach Sprache annotiert, daher ist auch keine genaue Angabe zur Zahl der althochdeutschen Texte möglich.) Bei „NA“ handelt es sich um nicht datierte Texte.

**3** Öffentlich verfügbare Version, d.h. ohne das Teilkorpus Briefe.

**4** Aus Gründen der Einfachheit wird das DTA in diesem Aufsatz als Korpus bezeichnet und auch wie eines behandelt. Als unausgewogene Volltextsammlung unterscheidet es sich aber natürlich klar von den übrigen Korpora (mit Ausnahme des Altdeutschkorpus, dessen Zusammenstellung der Überlieferungslage geschuldet ist).

**5** Texte, für die das nicht möglich ist, bleiben unberücksichtigt. Das betrifft im Altdeutschkorpus und in ReM Texte, die nur auf das Jahrhundert genau eingeordnet werden können oder für die gar keine zeitliche Einordnung möglich ist.

**6** Mit zunehmender Standardisierung und Loslösung von Autor- und Verlagsstandort verlieren sie zwar an Aussagekraft, für die älteren DTA-Texte sind die Angaben aber durchaus sinnvoll.

**Tabelle 1:** Zusammensetzung der diachronen Korpora des Deutschen.

| Korpus | Zeitschnitte | (Dialekt-)Raum | Textsorten/Textthemen |
|---|---|---|---|
| Referenzkorpus Altdeutsch | 50–100 Jahre, nicht balanciert | nicht balanciert | 6 |
| Referenzkorpus Mittelhochdeutsch | 50–100 Jahre, teilweise balanciert | teilweise balanciert | 6 |
| Bonner Frühneuhochdeutschkorpus | 50 Jahre[7], balanciert | 10, balanciert: Mittelbairisch, Schwäbisch, Ostfränkisch, Obersächsisch, Ripuarisch, Osthochalemannisch, Ostschwäbisch, Elsässisch, Hessisch, Thüringisch | 6, nicht balanciert |
| GerManC | 50 Jahre, balanciert | 5, balanciert: Ostoberdeutsch, Westoberdeutsch, Ostmitteldeutsch, Westmitteldeutsch, Niederdeutsch[8] | 7, balanciert (+1 Briefe, nicht öffentlich) |
| Deutsches Textarchiv | kontinuierlich, nicht balanciert | keine Gliederung | 4, nicht balanciert, mit zahlreichen Unterkategorien |
| DWDS-Kernkorpus 20, 21 | 10 Jahre, balanciert | keine Gliederung | 4, balanciert (DWDS 20) bzw. 1 (DWDS 21) |

Textsorte oder Hauptthema eines Textes werden in jedem Korpus anders definiert und kategorisiert. Hier wurde die Kategorie „Topic“ aus dem Altdeutschkorpus und ReM gewählt und auf die übrigen Korpora (außer DWDS-Kernkorpora) übertragen (Tabelle 2). Wie ersichtlich wird, ist der Bereich „Alltag“ über die Korpora hinweg sehr heterogen, in den anderen Bereichen ist mit größerer Kontinuität zu rechnen. Für textsortensensible Untersuchungen sollten sie bevorzugt werden.[9]

**7** Mit internen Lücken von 50 Jahren (d.h. keine Texte für 1400–1449, 1500–1549, 1600–1649).
**8** Hochdeutsche Texte aus dem niederdeutschen Raum.
**9** Die Kategorisierung ist sehr grob. Sie wird hier nur genutzt, um die Korpuszusammensetzungen vergleichbar zu machen, nicht aber für die Fallstudien.

**Tabelle 2:** Zuordnung von Textsorten/-themen auf Basis der Kategorie „Topic“ aus Altdeutschkorpus und ReM. Im DTA sind darüber hinaus einige unkategorisierte Texte vorhanden.

| AD | ReM[10] | Bonner | GerManC | DTA |
|---|---|---|---|---|
| Alltag | Alltag | chronikalische und Berichtstexte | Zeitungen | Überkategorien Zeitung, Gebrauchsliteratur (außer Unterkategorien Amtsdruckschrift, Kolportageliteratur, Recht, Verordnung, Bibelübersetzung, Erbauungsliteratur, Leichenpredigt, Theologie, Astrologie, Astronomie, Lexikon, Naturwissenschaft, Ökonomie, Pädagogik, Philosophie, Poetik, Pschologie, Zoologie) |
| Literatur | Literatur | unterhaltsame Texte | Erzählprosa, Dramen | Überkategorie Belletristik (außer Unterkategorien Lyrik, Reimpaarspruch, Religiöse Reimpaarerzählung, Schäferdichtung) |
| Poesie | Poesie | NA | NA | Unterkategorien Lyrik, Reimpaarspruch, Schäferdichtung |
| Recht | Recht | Rechts- und Geschäftstexte | Rechtstexte | Unterkategorien Gelegenheitsschrift: Vertrag, Jura, Recht, Verordnung |
| Religion | Religion | Bibeltext, erbauliche Texte, kirchl.-theol. Fachtexte | Predigten | Unterkategorien Bibelübersetzung, Erbauungsliteratur, Leichenpredigt, Ordensliteratur, Religiöse Reimpaarerzählung, Theologie |
| Wissenschaft | Wissen (-schaft) | | (Natur-)Wissenschaften, Geisteswissenschaften | Oberkategorie Wissenschaft (außer Unterkategorien Jura, Recht, Theologie) |

**10** Bei Mehrfachzuordnungen (z. B. „Religion, Alltag“) Zuordnung zum ersten Topic.

Ein Vergleich nach Dialektraum und Textthema zeigt starke Divergenzen zwischen Altdeutschkorpus, ReM, Bonner Korpus und GerManC (Abbildung 3–6).[11] Eine einheitliche Umkategorisierung wie die hier vorgeschlagene ist für eine statistische Analyse dieser Faktoren unerlässlich.[12]

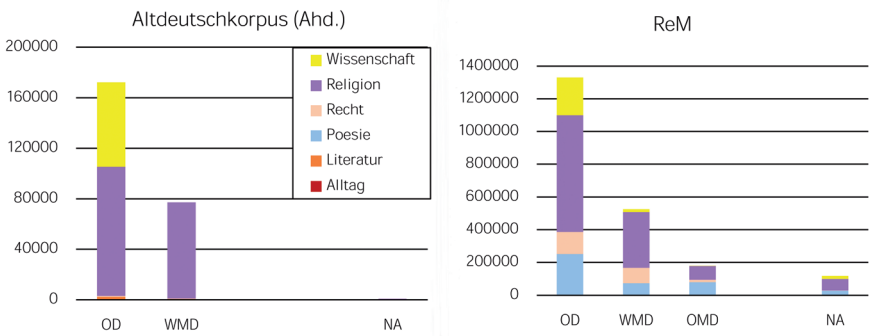



**Abbildung 3 und 4:** Dialektraum und Textthema in Altdeutschkorpus (nur Ahd.) und ReM. (OD=Oberdeutsch, WMD=Westmitteldeutsch, OMD=Ostmitteldeutsch, ND=hochdeutsche Texte des niederdeutschen Sprachraums.)

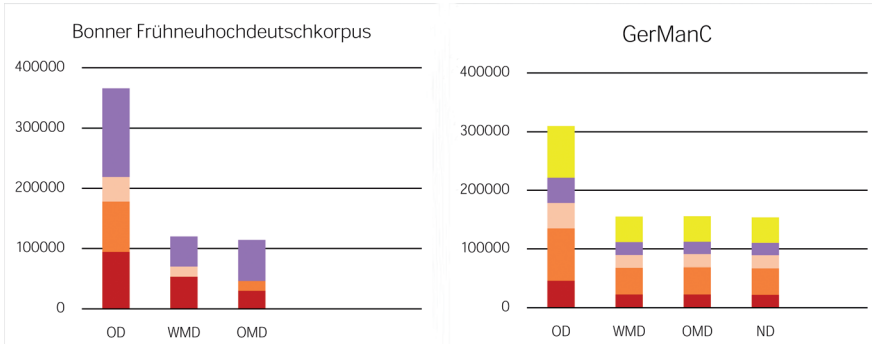



**Abbildung 5 und 6:** Dialektraum und Textthema in Bonner Frühneuhochdeutschkorpus und GerManC.

11 Für das DTA werden keine Visualisierungen präsentiert, da der abgedeckte Zeitraum so groß ist, dass eine Gesamtverteilung auf Textsorten und Räume wenig Aussagekraft hätte.

12 Für die hier vorgestellten Untersuchungen wurde auf statistische Verfahren verzichtet, da die damit einhergehenden methodischen Probleme und Überlegungen eine umfassendere Dokumentation erfordern als der gegebene Rahmen ermöglicht.

## 3.3 Tagging, Annotation, Abfragemöglichkeiten und Export

Alle Korpora sind lemmatisiert und wortartengetaggt, das Bonner Frühneuhochdeutschkorpus allerdings stark eingeschränkt (nur Substantive, Adjektive und Verben, andere Wörter sind lediglich über die Wortform abfragbar). Altdeutschkorpus, ReM, Bonner Korpus und GerManC verfügen für die getaggten Wörter auch über morphologische Informationen, GerManC weist außerdem Dependenzparsing und grammatische Funktionen auf, vgl. Tabelle 3.

**Tabelle 3:** Tagging und linguistische Annotation der diachronen Korpora.

| Korpus | Tagset | linguistische Annotation | Dokumentation |
|---|---|---|---|
| Referenzkorpus Altdeutsch | DDDTS | morphologische Informationen<br>Satztyp (linear) | Donhauser et al. (2018) |
| Referenzkorpus Mittelhochdeutsch | HiTS | morphologische Informationen | Klein & Dipper (2016) |
| Bonner Frühneuhochdeutschkorpus | keines; Unterscheidung von Substantiven, Adjektiven & Verben | morphologische Informationen (nur Substantive, Adjektive, Verben) | Fisseni (2017) |
| GerManC | STTS-EMG | morphologische Informationen<br>Dependenzrelationen[13] | Durrell et al. (2012) |
| Deutsches Textarchiv | STTS | – | Haaf et al. (2018) |
| DWDS-Kernkorpus 20, 21 | STTS | – | Geyken (2007) |

Altdeutschkorpus, ReM und Bonner Frühneuhochdeutschkorpus sind über ANNIS abfragbar, GerManC via COSMAS II (allerdings ohne die Annotationen, hierfür ist eine eigene Installation von z.B. GATE nötig), DTA und DWDS über die Suchmaske des DWDS (und DTA zusätzlich auch auf der eigenen DTA-Seite). Alle Korpora erlauben den Export von Beleg und Kontext, Annotationen und Lemmata sind allerdings nicht in derselben Zeile (Altdeutschkorpus, ReM, Bonner) bzw. gar nicht (GerManC, DTA, DWDS) exportierbar, was ein Problem für die Nachannotation und Auswertung darstellt. Aus diesem Grund und wegen der höheren Ge-

**13** Die Annotationen wurden mit einem am Nhd. trainierten Tool automatisiert vorgenommen und sind entsprechend nur sehr eingeschränkt nutzbar (Durrell et al. 2012: 11).

schwindigkeit wurden die Korpora, wo möglich, über eine lokale Installation der Corpus Workbench (CWB; Evert & Hardie 2011) abgefragt, die zeilenweisen Export von Belegstelle, Lemma, beliebigen Annotationen und Metadaten erlaubt.[14] Lediglich für Altdeutschkorpus, DTA und DWDS wurden die Online-Suchoberflächen genutzt. Die CWB-Abfragen sind jedoch online exakt reproduzierbar.

# 4 Testfälle

Im Folgenden werden vier Fallstudien vorgestellt, die sich dazu eignen, die Vergleichbarkeit von Bonner Korpus, GerManC und DTA zu prüfen. Ausgewählt wurden Phänomene, die möglichst wenig auf korpusspezifische Datenaufbereitung angewiesen sind, d.h. nicht auf Wortartinformationen oder komplexe Kombinationen verschiedener Kriterien zurückgreifen. Es handelt sich um zwei graphematische und zwei lexikalische Fallstudien. Für die Zukunft ist eine Ergänzung durch weitere Untersuchungen auf anderen linguistischen Ebenen wünschenswert.

## 4.1 Fallstudie 1: Virgelgebrauch

### 4.1.1 Forschungsstand

Ab dem frühen 16. Jh. verbreitet sich die Virgel zur internen Gliederung von Sätzen in Drucken mit gebrochenen Schriften (Masalon 2014: 35). Ihr steht zur gleichen Zeit das Komma in Antiquaschriften gegenüber; Ebert et al. (1993: 29–30) sprechen von der Virgel als „Langform des Kommas“ und vom Komma als „Kleinform“ der Virgel. Diese komplementär distribuierte Allographie entwickelt sich in der Folge in gebrochenen Schriften zu freier Variation zwischen Komma und Virgel (während sich umgekehrt keine Ausbreitung der Virgel in Antiquaschriften beobachten lässt). Schließlich kommt es im 18. Jh. zu einem vollständigen Abbau der Virgel (Ebert et al. 1993: 30).[15]

Die fnhd. Virgel ist jedoch nicht ausschließlich als Entsprechung des Kommas zu betrachten, sie „konkurriert mit den meisten anderen Zeichen (bes. dem Punkt) in einigen […] Funktionen (Redeschluss, Satz- bzw. Redegliederung)“

**14** Für die CWB-aufbereiteten Korpusdateien danke ich Susanne Flach (Neuchâtel) und Stefan Hartmann (Bamberg). Die Open-Source-Software ist über http://cwb.sourceforge.net/ beziehbar.
**15** Als satzzeichenähnliche Markierung tritt sie heute noch bei der Wiedergabe von Verstexten anstelle eines Zeilenumbruchs auf.

(Ebert et al. 1993: 29). Auch die Anfangsgroßschreibung von Teilsätzen lässt sich als Entsprechung (und Vorläufer) der Virgel betrachtet. Die Korpusdaten von Masalon (2014: 247–273) zeigen maximale Virgelanteile im 16. Jh. (81 %) und einen drastischen Rückgang zum 18. Jh. hin (10,9 %), ab dem 19. Jh. gibt es keine Belege mehr (vgl. gepunktet Linie in Abbildung 7).[16]

### 4.1.2 Studie

Da der Virgelgebrauch v.a. in Drucken zu erwarten ist und das Zeichen im 19. Jh. nicht mehr auftritt, werden muss die Fallstudie auf Bonner Frühneuhochdeutschkorpus, GerManC und DTA beschränkt bleiben. Die Satzzeichen des Bonner Korpus sind über das Annotationsmerkmal „zeichen“ abrufbar, das als Eigenschaft des ihm vorangehenden Wortes definiert ist. In den anderen beiden Korpora kann eine normale Tokensuche genutzt werden.

Aufgrund der Fragestellung müssen aus dem Bonner Korpus alle Texte ausgeschlossen werden, bei denen es sich um Handschriften (14) und/oder Editionen (14) handelt, damit entfällt der erste Zeitschnitt vollständig.[17] Grund ist, dass die Virgel eng an ihr Produktionsmedium und die verwendete Schrift gebunden ist und überdies häufig auch in sorgfältigen Editionen die Zeichensetzung nicht mit dem Original übereinstimmt.[18] Eine zuverlässige Untersuchung des Virgelgebrauchs muss sich auf Korpora mit Originaldrucken als Grundlage beschränken.

Trotz dieser Homogenisierung der Korpusgrundlage zeigt sich eine deutliche Divergenz zwischen Bonner Korpus und DTA zu Beginn des Untersuchungszeitraums (Abbildung 7).

**16** Verglichen mit Punkt, Doppelpunkt, Komma, Semikolon, Ausrufezeichen und Fragezeichen. Ausgewertet wurden pro Zeitschnitt 50 Textausschnitte à 500 Wörter.

**17** Text-IDs Handschriften: 111, 121, 131, 141, 151, 211, 221, 231, 241, 251, 113, 133, 213, 253; Editionen: 111, 113, 123, 133, 141, 151, 211, 213, 221, 231, 233, 235, 237, 241, 251, 253. Es verbleiben in Zeitschnitt 2 6 Texte, in 3 und 4 je 9 Texte.

**18** In keiner der 14 Hss. treten Virgeln auf, auch nicht in den beiden, die keine Editionen zur Grundlage haben. In zwei der vier Druckeditionen werden Virgeln verwendet, in zweien nicht.

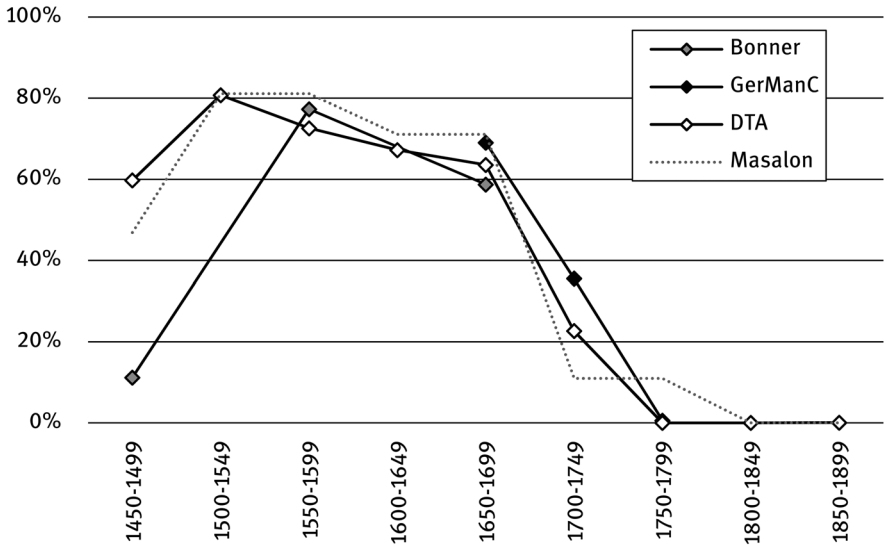



**Abbildung 7:** Virgelanteil im Verhältnis zu anderen Satzzeichen (Punkt, Doppelpunkt, Komma, Semikolon, Ausrufezeichen, Fragezeichen).[19] Daten aus Masalon (2014) zum Vergleich. (Da nach Jh. erhoben, nicht in 50-Jahres-Schritten, Doppeleintragung bei beiden Fünfzigjahresschritten.)

Eine genauere Analyse der jeweils enthaltenen Texte zeigt, dass die Varianz auch innerhalb der beiden Korpora besteht (Tabelle 4). Die Variationsbreite kann von der geringen Textanzahl nicht aufgefangen werden, weshalb der erste Zeitschnitt in keinem der beiden Korpora zuverlässig ist. Die DTA-Daten liegen zwar näher an Masalon (2014), der eine bessere Datengrundlage hat; das liegt jedoch nicht daran, dass das DTA ein realistischeres Abbild der typografischen Praxis seiner Zeit wäre als das Bonner Korpus, sondern ist reiner Zufall.

**Tabelle 4:** Anteile der Virgel an allen Satzzeichen für den Zeitschnitt 1450–1499 im DTA (grau) und Bonner Korpus (weiß).

| Text | Virgelanteil | Satzzeichen abs. | Anmerkung |
|---|---|---|---|
| nn_gottfried_1497 | 0 % | 2 | Verstext |
| oa_morgener_1497 | 0 % | 108 | Verstext |
| nn_almanach04_1492 | 0 % | 5 | Almanach (v.a. Listenformat) |

19 Die Daten wurden nicht nachträglich bereinigt, d.h. es sind – im Gegensatz zu Masalon (2014) – auch Ordinalzahlpunkte, Abkürzungspunkte, Bruchstriche etc. enthalten. Die DTA-Daten wurden bereits im November 2017 erhoben.

**Tabelle 4:** (fortgesetzt)

| Text | Virgelanteil | Satzzeichen abs. | Anmerkung |
|---|---|---|---|
| nn_almanach05_1473 | 0 % | 34 | Almanach (v.a. Listenformat) |
| 243 | 0 % | 957 | |
| nn_almanach05_1481 | 0 % | 128 | Almanach (v.a. Listenformat) |
| 143 | 1 % | 1.124 | |
| nn_dracole_1488 | 2 % | 107 | |
| 153 | 6 % | 1.066 | |
| nn_almanach04_1487 | 14 % | 14 | Almanach (v.a. Listenformat) |
| nn_almanach10_1493 | 19 % | 231 | Almanach (v.a. Listenformat) |
| 223 | 37 % | 1.403 | |
| nn_fusspfad_1492 | 85 % | 1.254 | |

Ab dem 16. Jh. sind die Ergebnisse gut mit Masalons (2014) Zahlen vergleichbar und ähneln sich untereinander deutlich: Die vier Überschneidungszeiträume weisen zwar Abweichungen auf (maximal 12 Prozentpunkte; zwischen Bonner und DTA 1700–1749), der Trend ist aber klar. Die Ähnlichkeiten bezüglich des Virgelgebrauchs können als erster Hinweis auf eine generelle Kombinierbarkeit der Korpora dienen.

## 4.2 Fallstudie 2: /aɪ̯/-Schreibung

### 4.2.1 Forschungsstand

Zwischen dem 14. und 16. Jh. variieren <i>, <j> und <y> wortmedial (teilweise auch initial) frei; ab Ende des 16. Jhs. geht der Gebrauch von <y> u.a. durch den Einfluss der Fruchtbringenden Gesellschaft zurück (Ebert et al. 1993: 43–44). Bei der Schreibung des Diphthongs /aɪ̯/ findet sich entsprechend <ey>, <ay> neben <ei>, <ai>. (Zur regional und konfessionell geprägten Wahl des ersten Diphthongbestandteils z.B. Schmid 1998: 18–19.) Heute hat sich die Schreibung <ei> durchgesetzt (mit vereinzelten <ai>-Schreibungen zur Homonymendifferenzierung, vgl. <Leib>, <Laib>).

### 4.2.2 Studie

Die Durchsetzung von <ei> gegenüber den anderen Graphien sollte sich also ebenfalls dazu eignen, zu bestimmen, ob und wie sehr die diachronen Korpora vergleichbar sind. Auch hier ist kein Rückgriff auf Annotationen nötig, die Diphthongschreibungen können mit regulären Ausdrücken gesucht werden.

Tatsächlich zeigt sich eine sehr große Übereinstimmung zwischen GerManC und DTA, zwischen Bonner Korpus und DTA kommt es jedoch zu einigen Abweichungen (Abbildung 8). Für den ersten DTA-Zeitraum ist das nicht verwunderlich, denn wie bereits festgestellt, sind hier nur wenige verschiedene Texte vorhanden. Der zweite Zeitraum umfasst zwar mehr Texte (14), allerdings ist einer davon Luthers Septembertestament mit 97,8 % <ey>-Schreibungen. Da es durch seinen Umfang 60,3 % der untersuchten Diphthongschreibungen in diesem Zeitraum beisteuert (und zudem noch drei weitere, kürzere Texte ebenfalls von Luther stammen), wird das Ergebnis massiv verzerrt. Das ist auch für andere Untersuchungen in diesem Zeitraum zu befürchten: Es wird nicht die Sprache der ersten Hälfte des 16. Jhs. untersucht, sondern die Sprache Luthers.[20]

Soll der Zeitraum 1500–1549 in einer Korpusuntersuchung berücksichtigt werden, so muss den Lutherdaten besondere Aufmerksamkeit geschenkt werden. In vielen Fällen wird es sich empfehlen, sie vollständig herauszunehmen. Ob der verbleibende Textumfang für die vorliegende Fragestellung geeignet ist, muss dann überprüft werden. In späteren Zeitschnitten fallen umfangreichere Werke in den meisten Fällen nicht mehr besonders ins Gewicht.

**20** Ebert et al. (1993: 44) beobachten bei Luther freie Variation zwischen <i> und <y>, mit einer Vorliebe für <y>, allerdings weitaus weniger ausgeprägt als im Septembertestament („Bach 1974 bucht bis 1520 einen Anteil von rund 25 % y-Belegen bei Luther"). Dass Luther selbst (und nicht nur seine Drucker) die <ey>-Schreibung stark bevorzugte, zeigt der Vergleich von Manuskripten und Drucken bei Haubold (1914: 72–73): Bei den 76 Belegen, die sich in dieser Hinsicht zwischen Hs. und Druck unterscheiden, hat der Druck in 48 Fällen <i> wo Luther <y> hat, in nur 28 Fällen ist es umgekehrt. (Für <ey> alleine sind es 35 bzw. 12; Übereinstimmungen zwischen Hs. und Druck wurden nicht erfasst.)

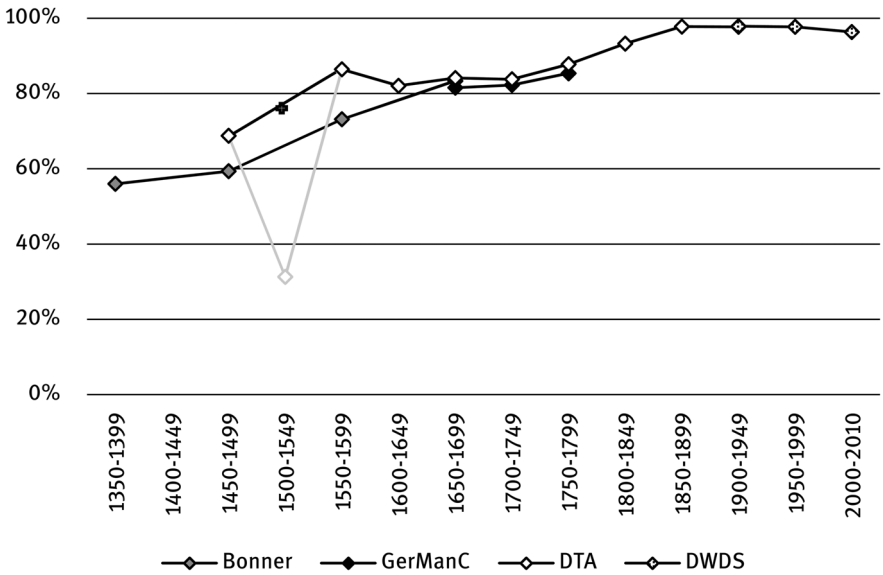



**Abbildung 8:** Anteil an <ei>-Schreibungen für den Diphthong /a͡ɪ/ (ggü. <ey>, <ay>, <ai>) (Bonner n=45.384; GerManC n=71.254; DTA n=18.421.675; DWDS n=12.225.934). Das schwarze Kreuzchen zeigt die DTA-Daten ohne das Septembertestament an, die ausgegraute Raute mit.

## 4.3 Fallstudie 3/4: Lexemfrequenzen: *Gott*, *und*

### 4.3.1 Forschungsstand

Die letzten Teilstudien stützen sich nicht auf konkrete Forschungsergebnisse zum Thema, sondern leiten die Erwartung für das Verhalten im Korpus etwas indirekter ab. Für die Lexeme *Gott* und *und* wird mit einem diachronen Rückgang der Gebrauchsfrequenz gerechnet. Bei *Gott* wird die zunehmende Ausdifferenzierung von Textsorten im Frühneuhochdeutschen, mit der verminderte Gottesbezüge einhergehen, als Grund vermutet. Eine syntaktische Ausdifferenzierung liegt bei *und* vor: Mit der Grammatikalisierung neuer satzverknüpfender Mittel und dem fortschreitenden Ausbau hypotaktischer Strukturen ist davon auszugehen, dass die Gebrauchsfrequenz der Konjunktion sinkt.

Da keine Vergleichsdaten vorliegen, sind Zu- oder Abnahme in den vorliegenden Studien allerdings nicht besonders aussagekräftig, der Fokus liegt hier auf der Analyse von Konvergenzen und Divergenzen zwischen den einzelnen Korpora.

### 4.3.2 Studie

Lexikalische Veränderungen zur Überprüfung der Korpora bieten sich neben graphematischen besonders an, weil sie höchstens auf die Lemmatisierung zugreifen. Insbesondere bei maschinell einfach erkennbaren Lexemen ist mit wenig Verzerrung zu rechnen (anders aber in Kap. 5.1). In der vorliegenden Untersuchung wurde *Gott* über die Lemmasuche erhoben, *und* über einen regulären Ausdruck, der Schreibvarianten (<v->, <-nn->, <-t>) berücksichtigt. Letzteres war nötig, weil das Bonner Korpus keine Lemmatisierung für die Wortart besitzt. Wichtig war bei der Auswahl der Lexeme, dass insgesamt mit einer hohen Frequenz in den Korpora gerechnet werden kann, weil dahinter ein kulturell salientes Konzept steht (*Gott*) bzw. es sich um ein Funktionswort handelt (*und*).

Tatsächlich zeigen alle Korpora für *Gott* einen Rückgang (Abbildung 9).[21] Bei *und* setzt der Rückgang erst Anfang des 17. Jhs. ein, innerhalb des Bonner Korpus zeigt sich kein klarer Trend (Abbildung 10).

Die Überschneidungszeiträume zeigen größere Divergenz zwischen den Korpora zu Beginn, ab der zweiten Hälfte des 17. Jhs. haben sich die Werte aber deutlich angeglichen, was als erster Hinweis auf Vergleichbarkeit gesehen werden kann. Ob und welche Lexeme tatsächlich als zuverlässiges Maß für die Korpusvergleichbarkeit dienen können, kann diese erste Probebohrung allerdings noch nicht beantworten.

**21** Ursprünglich wurden für *Gott* auch ReM-Daten erhoben, die hier nicht einbezogen werden, um den Untersuchungsrahmen konstant zu halten. Im ReM-Zeitraum zeigt sich keine klare Zu- oder Abnahme, die Daten haben einen leicht wellenförmigen Verlauf.

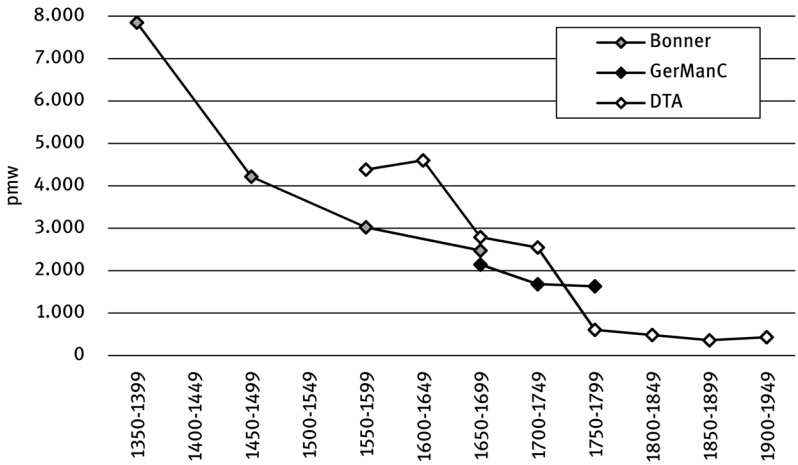



**Abbildung 9:** Gebrauchsfrequenz des Lexems *Gott* pro Million Wörter (Bonner n=2.587, GerManC n=1.409, DTA n=282.248).[22]

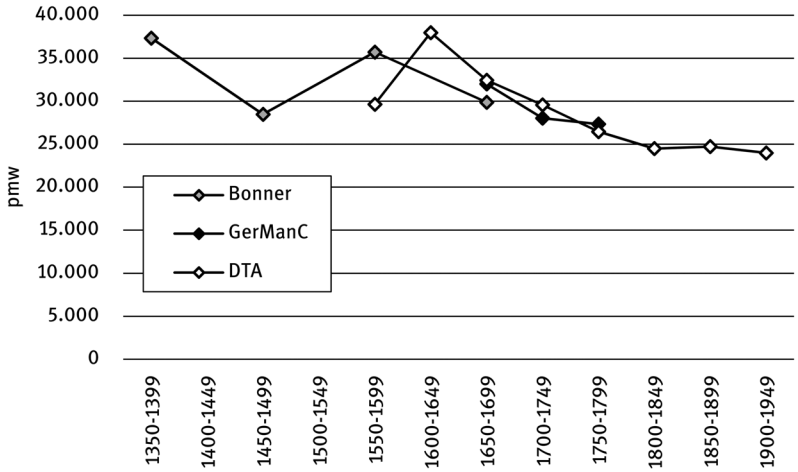



**Abbildung 10:** Gebrauchsfrequenz des Lexems *und* pro Million Wörter (Bonner n=19.603, GerManC n=22.555, DTA n= 5.792.319).

**22** Für das DTA wurden die unzuverlässigen Zeitschnitte 1450–1499 und 1500–1549 nicht berücksichtigt. Beim zweiten Zeitschnitt geschieht das, statt dass die Werte des Septembertestaments herausgerechnet werden (wie es bei den prozentual berichteten <ei>-Anteilen der Fall ist), weil eine Neuberechnung der Korpusgröße ohne das Septembertestament derzeit mit erheblichem Aufwand verbunden wäre.

# 5 Weiterführende Untersuchungen

Im Folgenden kommen durch komplexere Abfragen zusätzliche potenzielle Verzerrungsfaktoren ins Spiel: Zum einen kann es Probleme mit der Lemmatisierungs- und Annotationsqualität der Korpora geben, zum anderen stellt es hier eine Herausforderung dar, Suchabfragen zu konstruieren, die über die Korpora hinweg vergleichbar sind. Bei einer der Studien (Kap 5.1) sind die Belegzahlen außerdem relativ niedrig. Es handelt sich also um Untersuchungen, die der linguistischen Realität wesentlich mehr entsprechen als die Fingerübungen aus Kap. 4.[23]

## 5.1 Fallstudie 4: Stammerweiterung schwacher Maskulina

### 5.1.1 Forschungsstand

In (früh-)neuhochdeutscher Zeit ist bei zahlreichen unbelebten, ehemals schwachen Maskulina ein sukzessiver Umbau des Paradigmas zu beobachten (Nübling 2008: 305–308).[24] Dabei entsteht zunächst durch einen schwach-starken Stapelgenitiv (-*n*-*s*) eine Sondergruppe (in Abbildung 12 als „gemischt“ bezeichnet). Daran schließt sich eine *n*-Erweiterung im Nominativ Singular an, was zu einer Reanalyse des Stamms in allen Kasus führt, sodass nun, wie für starke Maskulina auf Reduktionssilbe üblich, nur der Genitiv Singular eine eigenständige Endung aufweist. Im letzten Schritt kann im Plural ein Umlaut hinzukommen.

**23** Ursprünglich wurden auch Daten für eine dritte komplexe Studie erhoben, zum Gebrauch des Dativ-*e*. Zu Abfrageproblemen (die Studie musste auf ein Einzellexem, *Haus*, eingeschränkt werden) kommt hier die bekanntermaßen wechselhafte Geschichte hinzu (Abbau, teilweise Restitution, erneuter Abbau). Diese Aspekte können hier leider nicht diskutiert werden.

**24** Die verbleibenden schwachen Maskulina gehören in der Folge einem phonologisch und semantisch konditionierten Schema an, vgl. Köpcke (1995).

| | Mhd. > | Frühnhd. > | Nhd. > | heute |
|---|---|---|---|---|
| | schwach [–belebt] | gemischt → *s*-Genitiv | stark (–UL) → *n*-Erweiterung im Nom. Sg. | stark (+UL) → Umlaut im Pl. |
| N | der schade | der schade | der Schade**n** | der Schaden |
| G | des schade-n | des schade-n**s** | des Schaden-s | des Schaden-s |
| D | dem schade-n | dem schade-n | dem Schaden | dem Schaden |
| A | den schade-n | den schade-n | den Schaden | den Schaden |
| N–A | die schade-n | die schade-n | die Schaden-Ø | die Schäden-Ø |
| | Klasse wird geräumt; reduziert sich auf +belebt: *Affe*, *Kunde*, *Matrose*, *Geselle* | *Name*, *Buchstabe*, *Gedanke* | *Balken*, *Brunnen*, *Knochen* | *Schaden*, *Garten*, *Magen* |
| | Übergänger: (Schwankungsfälle) | | *Funke/n* *Glaube/n* | *Bogen* *Wagen* |

**Abbildung 11:** Räumung der schwachen Maskulina (Abbildung aus Nübling 2008: 307).

Für 11 Substantive liegen bei Kopf (2018: 410) Korpusdaten zum Anteil des erweiterten Genitivs in Bonner Korpus und DTA vor. Er dominiert schon Anfang des 17. Jhs.[25] Damit dürfte auch die Stammerweiterung des Nominativs in den Überschneidungsbereich der drei Korpora fallen.

### 5.1.2 Studie

Die korpuslinguistische Bestimmung der Klassenzugehörigkeit in dieser Umbruchphase erfordert vollständige Paradigmen, idealerweise nicht nur innerhalb eines Zeitschnitts, sondern bei einzelnen Autorinnen und Autoren. Auf die damit verbundenen Schwierigkeiten hat u.a. Klein (2018) hingewiesen. Im Folgenden steht jedoch nicht das Gesamtparadigma im Zentrum des Interesses, sondern

25 Erste Hälfte des 17. Jhs., ausschließlich *ns*-Genitiv: *Glaube*, *Same*. Deutlich überwiegend *ns*-Genitiv: *Buchstabe*, *Friede*, *Name*, *Schatte*, *Wille*. Knapp überwiegend *ns*-Genitiv: *Brunne*. Zu wenige Daten: *Funke*, *Gedanke*, *Schlitte*.

Zeitpunkt und Ausbreitung der *n*-Erweiterung. Die Untersuchung zur Genitivform bei Kopf (2018: 410) zeigt zwar vereinzelt lexemspezifische Unterschiede, aber einen klaren gemeinsamen Trend, sodass es legitim erscheint, auch für die Nominativform mehrere Lexeme (*Brunnen*, *Haufen*, *Schatten*) gemeinsam auszuwerten. Alle drei sind heute keine Schwankungsfälle mehr und haben im Standard den Schritt zum Umlautplural nicht vollzogen.

Im Gegensatz zu einem Lexem wie *Gott* kann man sich bei der Abfrage von Lemmata, deren Nennform im Untersuchungszeitraum variiert, und die zudem bei Kleinschreibung strukturell wie Infinitive aussehen, nicht immer auf die korpuseigene Lemmatisierung verlassen. Insbesondere die Lemmatisierungsqualität des DTA erweist sich bei schwankender Nennform als hochproblematisch. Die Gegenüberstellung einer Wortformsuche (mit regulären Ausdrücken, Details s.u.) und einer Lemmasuche am Beispiel *Brunne*(*n*) liefert den folgenden Befund: Über den gesamten Untersuchungszeitraum machen *n*-lose Formen bei der Wortformensuche 8,5 % aus, bei der Lemmasuche dagegen nur 1,5 %. Acht Wortformen sind auf sechs Lemmata verteilt, wobei es kaum 1:1-Zuordnungen gibt (Abbildung 12).

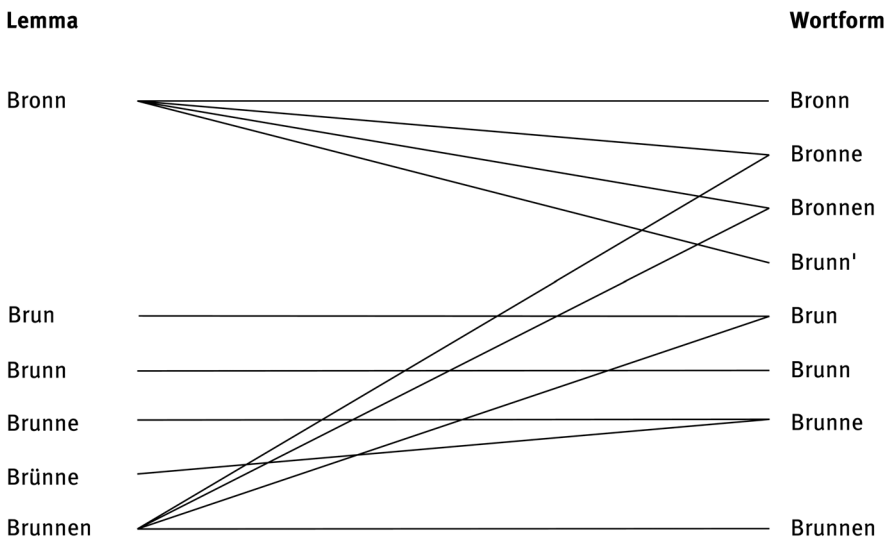



**Abbildung 12:** Verhältnis von Lemmata und Wortformen für das Lexem *Brunne(n)* im DTA (nur Nominativ Singular).

Daher wurden für GerManC und DTA mit regulären Ausdrücken potenzielle Schreibungen generiert, die idealerweise alle denkbaren Wortformen abdecken:

| | |
|---|---|
| *Brunne(n)* | (B\|b)r(u\|o)nn?e?n? |
| *Haufe(n)* | (H\|h)(a\|o)uff?e?n? |
| *Schatte(n)* | (S\|s)chatt?e?n? |

Um möglichst viele Nicht-Nominativ-Singulare auszuschließen, wurde dem Suchwort ein Definit- oder Indefinitartikel (*der*, *ein* in verschiedenen Schreibungen) vorangestellt. Das Bonner Korpus weist dagegen eine sehr sorgfältige Lemmatisierung auf, da es speziell zur Untersuchung von Flexion erstellt wurde. Hier konnte direkt auf die Formen aus der Lemmaliste zurückgegriffen und nach Treffern im Nominativ Singular gesucht werden, eine Einschränkung über Artikel war nicht nötig. Zwar hätte es theoretisch zu besserer Vergleichbarkeit der Ergebnisse geführt, wenn die Abfrage identisch gewesen wäre (also überall Suche mit Artikel), allerdings liefern die kleinen Korpora schon bei der feineren Suche so wenige Daten, dass eine weitere Beschneidung die Ergebnisse unbrauchbar gemacht hätte.

Die Kombination der Daten (Abbildung 13) zeigt 1550 und 1650 Schwankungen von bis zu 15 Prozentpunkten bei geringen Belegzahlen in GerManC und Bonner Korpus. Dennoch lässt sich der Beginn der *n*-Erweiterung auf die zweite Hälfte des 15. Jhs. datieren und eine kontinuierliche, nicht sprunghafte Ausbreitung annehmen. Hier zeigt sich nun die Stärke des kombinierenden Verfahrens: Die Beobachtungen auf Basis der kleinen Fallzahlen aus Bonner Korpus und GerManC werden durch die DTA-Daten gestützt.

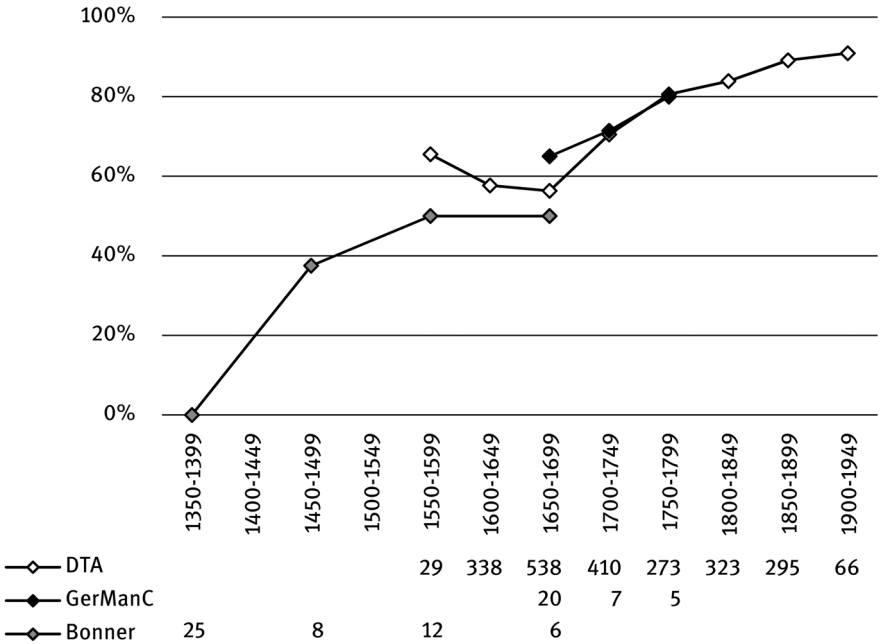



**Abbildung 13:** *n*-Erweiterungsanteile der Lexeme *Brunne(n)*, *Haufe(n)*, *Schatte(n)*. (Zahlen in der Tabelle: alle Treffer. Apokopierte Formen des Typs *Brunn* gelten als nicht erweitert.)

## 5.2 Fallstudie 5: Ausbreitung des Indefinitartikels

### 5.2.1 Forschungsstand

Während die Grammatikalisierung des Indefinitartikels *ein-* aus dem Zahlwort *ein* gesichert ist, fehlt nach wie vor eine umfassende empirische Untersuchung zu seiner funktionalen Veränderung und Ausbreitung. Szczepaniak (2016) rekonstruiert aus alt- und mittelhochdeutschen Belegen einen möglichen Grammatikalisierungsverlauf und betont die Notwendigkeit einer „systematischen Korpusuntersuchung". Dazu kann die vorliegende Studie möglicherweise methodische Anregungen beitragen.

Bereits im Althochdeutschen verändert sich der Gebrauch des Zahlworts (Oubouzar 1997: 170–174), der Großteil des Grammatikalisierungsprozesses fällt in mittelhochdeutsche Zeit. Aus inhaltlichen und methodischen Gründen wird die Ausbreitung des Indefinitartikels hier nur in einem bestimmten syntaktischen Kontext, nämlich in Präpositionalphrasen, untersucht (zur methodischen Begründung s. 5.2.2): Bei Präpositionalphrasen werden syntaktische Beziehungen durch

die Präposition geklärt, während eigenständige Nominalphrasen stärker auf kasusanzeigende Artikel angewiesen sind (Pavlov 1983: 36–37; vgl. auch Kopf 2018: 74–75). Daher ist damit zu rechnen, dass die Hauptzeit der Ausbreitung des Indefinitartikels in Präpositionalphrasen nach der Ausbreitung in Nominalphrasen liegt. Sie fällt voraussichtlich in die frühneuhochdeutsche Zeit und eignet sich daher für eine Untersuchung mit Bonner Korpus, GerManC und DTA.[26] Ein Vergleich des Indefinitartikels in NPs gegenüber PPs wird hier nicht geleistet, wäre aber von großem Interesse.

### 5.2.2 Studie

Die Einschränkung auf Präpositionalphrasen erfolgt nicht nur wegen des vermuteten Ausbreitungszeitraums, sondern auch wegen der guten Abgrenzbarkeit. Da der Gebrauch des Indefinitartikels mit seinem Nichtgebrauch kontrastiert werden muss,[27] um eine Ausbreitung beobachten zu können, ist eine Suchanfrage sinnvoll, bei der auch Nichtvorkommen klar erkennbar ist. Die Präposition begrenzt die Phrase klar und geht dem potenziellen Artikel direkt voraus, es kann also die erste Position nach der Präposition untersucht werden.

Die Erhebung wird auf die Präposition *mit* beschränkt, da 1. keine Verwechslungsgefahr mit homographen Formen besteht (im Gegensatz zu <in> ‚ihn' vs. ‚in'),[28] 2. es sich um eine sehr frequente Präposition handelt und 3. nicht mit Präposition-Artikel-Klisen zu rechnen ist. Präpositionalphrasen mit zusätzlichen Modifikatoren (Possessiva, Adjektive, Genitivattribute) bleiben unberücksichtigt. Da der Indefinitartikel mindestens ins Mittelhochdeutsche zurückreicht, werden für diese Untersuchung auch das Altdeutschkorpus und ReM einbezogen. Im Gegensatz zu den Fallstudien in Kap. 4 sind komplexere und damit fehleranfäl-

**26** Der Definitartikelgebrauch setzt außerdem in Präpositionalphrasen später ein als in einfachen Nominalphrasen (Oubouzar 1997: 170, Szczepaniak 2013: 104–105). Das wird damit begründet, dass hier i.d.R. adverbiale Verwendung vorliegt, für die Definitheit nicht relevant ist (z.B. *fon himile* ‚vom Himmel', *in lenzen* ‚im Frühling'; Oubouzar 1997: 170). Ist die Kategorie Definitheit hier aber generell nicht von Belang, so ist auch für den Indefinitartikel mit einer verzögerten Ausdehnung auf Präpositionalphrasen zu rechnen.

**27** In der Visualisierung wird der Indefinitartikel mit allen anderen Vorkommen, d.h. Artikellosigkeit und Definitartikelgebrauch, kontrastiert. Die Verschiebung im Definitheitsbereich (von Nullartikel zu Definitartikel) bleibt entsprechend unsichtbar. In der zweiten Hälfte des 18. Jahrhunderts ist im DTA mit rund 14 % Indefinitartikelgebrauch in *mit*-PPs der heutige Stand erreicht. (Definitartikel: 42 %, kein Artikel: 44 %.)

**28** Entsprechend schützt man sich vor falschen Lemmatisierungen, wie sie insbesondere im DTA auftreten können.

ligere Suchabfragen notwendig, die möglicherweise zu kleineren Abweichungen zwischen den Korpora führen:

**Tabelle 5:** Erhobene Konstruktionen zum Artikelgebrauch in einfachen *mit*-PPs nach Korpus.

| Korpus | Erhobene Konstruktionen | Erläuterung |
|---|---|---|
| Altdeutsch | *mit* (+ potenzieller Artikel) + Substantiv im Dat.Sg., nicht gefolgt von Possessivum oder Adjektiv | Wegen im Gang befindlicher Artikelgrammatikalisierung überschneiden sich Demonstrativa/Zahlwörter und Artikel im Tagset (definit: DD.* bzw. indefinit: DI.*). |
| ReM | *mit* (+potenzieller Artikel) + Substantiv im Dat.Sg., nicht gefolgt von Possessivum oder Adjektiv oder Wortform in Genitiv | |
| Bonner | *mit* (+ *d-*/*ein-*) + Substantiv im Dat. Sg., nicht gefolgt von Adjektiv oder Wortform in Genitiv | Nachgestellte Possessiva nicht mehr üblich und nicht getaggt, daher nicht explizit ausgeschlossen. Artikelwörter nicht getaggt, daher Wortformsuche mit Schreibvarianten. |
| GerManC | *mit* (+ Artikel) + Substantiv im Dat. Sg., nicht gefolgt von Adjektiv oder Wortform in Genitiv | Artikelgrammatikalisierung schon vollzogen, daher Suche nach Tag ART möglich. |
| DTA | *mit* (+ *d-*/*ein-*) + Substantiv | Keine Angaben zu Kasus und Numerus enthalten, daher nachträglich manuelle Entfernung von Pluralen und Phrasen mit Genitivattributen. |

Erste Vorkommen eines potenziellen Indefinitartikels sind in der 2. Hälfte des 12. Jhs. in ReM zu beobachten (Abbildung 14). Später steigt der Indefinitartikelgebrauch in den einzelnen Korpora deutlich, allerdings zu unterschiedlichen Zeiten. Der Gebrauch im Bonner Korpus schließt sich zunächst an den in ReM an, um in der zweiten Hälfte des 15. Jhs. erheblich zuzunehmen (von 5 % auf 11,5 %). DTA und GerManC zeigen sich konservativer. Sie erreichen bzw. übertreffen die frühen 11,5 % des Bonner Korpus erst Ende des 17. (DTA, 13,8 %) bzw. Anfang des 18. Jhs. (GerManC, 13,6 %).

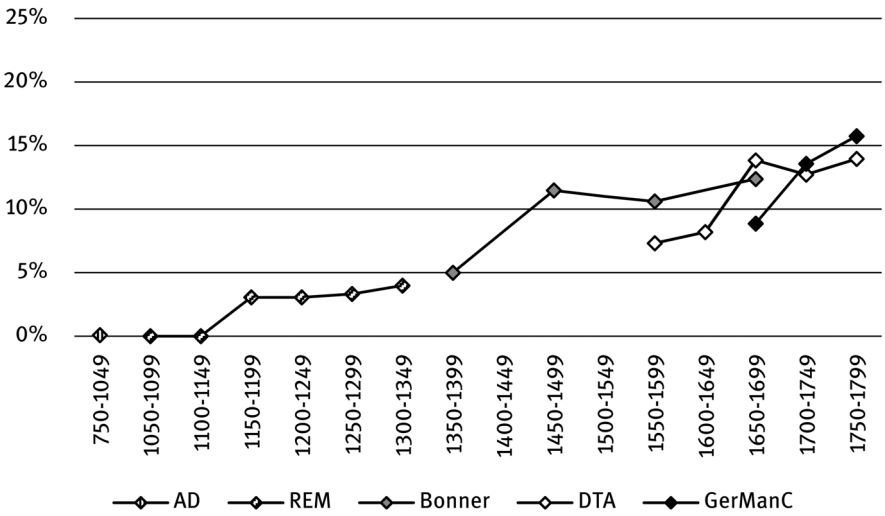



**Abbildung 14:** Anteile des Indefinitartikels in *mit*-PPs vs. artikellose PPs/PPs mit Definitartikel (AD n=1.108; ReM n=5.518; Bonner n=1.150; DTA n=4.774;[29] GerManC n=1.166). Skala bei 25 % abgeschnitten.

# 6 Fazit

Aus der Analyse der Korpora und den Fallstudien ergeben sich einige wichtige Punkte für den Umgang mit den Korpora. Für einzelne Zeitschnitte:

- Je nach Untersuchungsziel müssen im Bonner Korpus (1350–1399) Handschriften bzw. Editionen ausgeschlossen werden. (Liste s. Fn. 16, S. 11.)
- DTA-Zeitschnitt 1 (1473–1499) ist für die meisten sprachwissenschaftlichen Untersuchungen derzeit nicht geeignet (zu wenige Texte von sehr unterschiedlicher Länge).
- DTA-Zeitschnitt 2 (1500–1549) wird massiv von Luthers Septembertestament dominiert. Je nach Untersuchungsziel sollte er ausgeschlossen oder die Daten des Septembertestaments herausgerechnet werden.
- Insgesamt ist das DTA in den späteren Jahren (ab Mitte 17. Jhs.) zuverlässiger als zu Beginn.

**29** Aufgrund der notwendigen manuellen Nachbearbeitung wurde im DTA Zufallssamples von je 1.000 Belegen pro Zeitschnitt ausgewertet.

Für Abfragen:

- Suchabfragen sollten so einheitlich wie möglich gestaltet werden. Verfügt eines der Korpora nicht über Lemmatisierung, so sollte sie auch in den übrigen Korpora nicht genutzt werden, sondern ein einheitlicher regulärer Ausdruck für die Wortform.
- Wird Lemmatisierung genutzt, so muss insbesondere bei von Sprachwandel betroffener Nennform zunächst die Lemmatisierungsqualität sichergestellt werden.
- Bei niederfrequenten Phänomenen gehen bei einheitlicher einschränkender Abfrage in den kleinen, gut annotierten Korpora so viele Daten verloren, dass kein Vergleich mehr möglich ist. Hier muss auf die Annotation zurückgegriffen werden.
- Will man vergleichbare Datenqualität, so muss man ggf. mit einer offeneren Suche arbeiten (z.B. beim Indefinitartikel zwischen Präposition und Substantiv im Bonner Korpus einen Wortabstand von 0 bis 1 definieren, statt auf konkrete Artikelschreibungen einzuschränken) und nachträglich aussortieren.

Anhand der hier nicht vorgestellten Daten zum Dativ-*e* (vgl. Fn. 23, S. 18) wurde außerdem deutlich, dass z.B. regionale Ausgewogenheit von Korpora zwar wünschenswert ist, aber nicht genügt, wenn die Ergebnisse einer konkreten Suche nicht ebenfalls ausgewogen über die Regionen verteilt sind, sondern in verschiedenen Zeitschnitten jeweils verschiedene Regionen überwiegen. Für bekanntermaßen regionale oder textsortenspezifische Phänomene muss diese Ausgewogenheit nach der Erhebung statistisch hergestellt werden.

Insgesamt konnte gezeigt werden, dass die Kombination von Daten verschiedener Korpora in den Fallstudien zu gleichen Tendenzen führt. Das lässt für weitere, komplexere Untersuchungen Gutes hoffen. Für die Zukunft gilt es, einen Fundus weiterer Phänomene aufzubauen, deren Entwicklung bereits anhand anderer Daten als gesichert gelten kann, um mit einem großen, ausgewogener zusammengestellten Phänomenbündel zuverlässigere Einschätzungen treffen zu können.

# 7 Korpora

Bonner Frühneuhochdeutschkorpus. Universität Bonn; Werner Besch, Winfried Lenders, Hugo Moser & Hugo Stopp. Zugriff via https://korpora.zim.uni-duisburg-essen.de/annis/.

Deutsches Textarchiv (DTA). Berlin-Brandenburgische Akademie der Wissenschaften. Zugriff via www.dwds.de oder www.deutschestextarchiv.de.

Digitales Wörterbuch der deutschen Sprache (DWDS). Berlin-Brandenburgische Akademie der Wissenschaften. Zugriff via www.dwds.de.
GerManC. Universität Manchester, Martin Durrell et al. Download via http://ota.ox.ac.uk/desc/2544.
Referenzkorpus Altdeutsch. HU Berlin, Universität Frankfurt a.M., Universität Jena; Karin Donhauser, Jost Gippert, Rosemarie Lühr. Zugriff via http://www.deutschdiachrondigital.de.
Referenzkorpus Mittelhochdeutsch (ReM). Universitäten Bochum, Bonn; Klaus-Peter Wegera, Stefanie Dipper, Thomas Klein, Claudia Wich-Reif. Zugriff via www.linguistics.rub.de/rem/index.html.

# 8 Literatur

Christiansen, Mads (2016): *Von der Phonologie in die Morphologie.* Hildesheim, Zürich, New York: Olms.
Donhauser, Karin, Jost Gippert & Rosemarie Lühr (2018): *Referenzkorpus Altdeutsch. Dokumentation.* http://www.deutschdiachrondigital.de/dokumentation/manual/.
Durrell, Martin, Paul Bennett, Silke Scheible & Richard J. Whitt (2012): *The GerManC Corpus*. Manchester.
Ebert, Robert P. et al. (Hgg.) (1993): *Frühneuhochdeutsche Grammatik* (Sammlung kurzer Grammatiken germanischer Dialekte. A 12). Berlin: De Gruyter.
Evert, Stefan & Andrew Hardie (2011): Twenty-first century Corpus Workbench. Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*.
Fisseni, Bernhard (2017): *Das Bonner Frühneuhochdeutschkorpus* (FnhdC) 2017. https://korpora.zim.uni-duisburg-essen.de/FnhdC/Dokumentation.html.
Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In Christiane Fellbaum (Hg.), *Collocations and Idioms. Linguistic, lexicographic, and computational aspects*, 23–41. London: Bloomsbury.
Haaf, Susanne, Matthias Boenig, Christian Thomas & Frank Wiegand (2018): *Deutsches Textarchiv. Dokumentation*. http://www.deutschestextarchiv.de/doku.
Hartmann, Stefan (2016): *Wortbildungswandel. Eine diachrone Studie zu deutschen Nominalisierungsmustern*. Berlin, New York: De Gruyter.
Haubold, Fritz (1914): *Untersuchung über das Verhältnis der Originaldrucke der Wittenberger Hauptdrucker Lutherscher Schriften: Grunenberg, Lother, Döring-Cranach und Lufft zu Luthers Druckmanuskripten*. Borna-Leipzig: Noske.
Kempf, Luise (2016): *Adjektivsuffixe in Konkurrenz. Wortbildungswandel vom Frühneuhochdeutschen zum Neuhochdeutschen*. Berlin, New York: De Gruyter.
Klein, Andreas (2018): *Vom Beleg zum Paradigma. Empirische Probleme implikativer Klassenbestimmung* (Vortrag bei der Jahrestagung der Gesellschaft für germanistische Sprachgeschichte, 21.9.2018).
Klein, Thomas & Stefanie Dipper (2016): *Handbuch zum Referenzkorpus Mittelhochdeutsch* (Bochumer Linguistische Arbeitsberichte 19). Bochum.
Köpcke, Klaus-Michael (1995): Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache. *Zeitschrift für Sprachwissenschaft* 14(2), 159–180.

Kopf, Kristin (einger.): Was ist so besonders an Gott? Ein grammatischer Abweichler im Frühneuhochdeutschen. In Luise Kempf, Damaris Nübling & Mirjam Schmuck (Hgg.), *Linguistik der Eigennamen*.

Kopf, Kristin (2017): Fugenelement und Bindestrich in der Compositions-Fuge. Zur Herausbildung phonologischer und graphematischer Grenzmarkierungen in (früh)neuhochdeutschen N+N-Komposita. In Renata Szczepaniak, Nanna Fuhrhop & Karsten Schmidt (Hgg.), *Sichtbare und hörbare Morphologie*, 177–204. Berlin, New York: De Gruyter.

Kopf, Kristin (2018): *Fugenelemente diachron. Eine Korpusuntersuchung zu Entstehung und Ausbreitung der verfugenden N+N-Komposita*. Berlin: De Gruyter.

Masalon, Kevin C. (2014): *Die deutsche Zeichensetzung gestern, heute – und morgen (?): Eine korpusbasierte, diachrone Untersuchung der Interpunktion als Teil schriftsprachlichen Wandels im Spannungsfeld von Textpragmatik, System und Norm unter besonderer Berücksichtigung des Kommas*. Duisburg-Essen, Dissertation.

Nübling, Damaris (2008): Was tun mit Flexionsklassen? Deklinationsklassen und ihr Wandel im Deutschen und seinen Dialekten. *Zeitschrift für Dialektologie und Linguistik* 75(3), 282–330.

Oubouzar, Erika (1997): Zur Frage der Herausbildung eines bestimmten und eines unbestimmten Artikels im Althochdeutschen. *Cahiers d'études germaniques* 32, 161–175.

Pavlov, Vladimir M. (1983): *Von der Wortgruppe zur substantivischen Zusammensetzung* (Zur Ausbildung der Norm der deutschen Literatursprache (1470-1730) 4). Berlin: Akademie-Verlag.

Schmid, Hans U. (1998): Sprachlandschaften und Sprachausgleich in nachreformatorischer Zeit. Martin Luthers Bibelübersetzung in epigraphischen Zitaten des deutschen Sprachraums. *Zeitschrift für Dialektologie und Linguistik* 65(1), 1–41.

Szczepaniak, Renata (2013): *Grammatikalisierung im Deutschen: Eine Einführung*, 2. Aufl. Tübingen: Narr Francke Attempto.

Szczepaniak, Renata (2016): Vom Zahlwort eins zum Indefinitartikel ein(e). In Andreas Bittner & Klaus-Michael Köpcke (Hgg.), *Regularität und Irregularität in Phonologie und Morphologie*, 247–262. Berlin, Boston: De Gruyter.