4 Korpus

> Corpus linguistics studies languages on the basis of discourse. [...] It will never be possible to study all extant texts. All corpus linguistics can do is to work with a (suitable) sample of the discourse. Such a sample is called the corpus.
>
> (Teubert & Čermáková 2007: 41)

Die Verwendung von Korpora in Kritischen Diskursanalysen zielt oft darauf ab, über das Korpus quantitativen Zugang zum Diskurs zu erhalten. Für diesen Zweck muss das Korpus so gestaltet werden, dass es den zu betrachtenden Diskurs möglichst repräsentiert, wenn auch nie eine vollständige Übereinstimmung zwischen Diskurs und Korpus möglich ist. Im Rahmen dieser Arbeit wird der Diskurs über Klimawandel in Massenmedien der deutschsprachigen Schweiz betrachtet, insbesondere mit Fokus auf der Frage, wie Wissen aus Spezialdiskursen in massenmedialen Diskursen thematisiert und verhandelt wird. In diesem Sinne wird das Korpus so zusammengestellt, dass es sich möglichst an den hier beschriebenen Diskurs annähert.

4.1 Beschaffenheit

Das Korpus umfasst ausschliesslich Medien aus der deutschsprachigen Schweiz, die gemeinhin als Push-Medien verstanden werden, da Pull-Medien wie beispielsweise das Internet ein bereits bestehendes Interesse einer Person an einem bestimmten Thema sowie deren Bereitschaft, sich aktiv zu informieren, bedingen. Zudem repräsentieren öffentliche Push-Medien die vorhandenen Geltungsansprüche in der Gesellschaft (Tereick 2016: 28): «Medien sind Spiegel dessen, was gesellschaftlich anerkannt ist, wirken zugleich aber selbst wirklichkeitskonstitutiv.» Fernsehkanäle und Zeitungen scheinen unter diesen Medien aus rezeptiver Sicht von besondere Bedeutung zu sein, wie unter anderem Stamm, Clark & Reynolds Eblacas (2000) für den amerikanischen Raum aufzeigen; Parallelen zeigen sich auch für den Diskurs über Atomenergie in der Schweiz (Bonfadelli & Kristiansen 2012).

In das Korpus werden häufig rezipierte Push-Medien miteinbezogen, die einen möglichst hohen Anteil der Bevölkerung erreichen: Sämtliche Zeitungen, die 2014 eine Auflagenstärke von mehr als 100 000 Exemplaren aufwiesen (zehn Tageszeitungen, zwei Gratiszeitungen und sechs regionale Zeitungen; Auflagenstärken gemäss der WEMF AG für Medienforschung (WEMF 2014))[1], sowie die

1 Die Auflagenstärke entspricht der *total verbreiteten Auflage* in WEMF (2014). Es wurden lediglich deutschsprachige Zeitungen aufgenommen.

https://doi.org/10.1515/9783110760804-004

beiden Nachrichtenfernsehformate *10vor10* und *Tagesschau* des öffentlich-rechtlichen Fernsehens finden Eingang in das Korpus.[2] Der Zeitraum, über den Artikel in das Korpus aufgenommen wurden, reicht von 2007 bis 2014. Im Jahr 2007 wurde der vierte (Solomo et al. 2007; Parry et al. 2007; Metz et al. 2007), 2013 und 2014 der fünfte Sachstandsbericht des IPCC (Stocker et al. 2013; Field et al. 2014; Barros et al. 2014; Edenhofer et al. 2014) veröffentlicht. Grund für den Zeitraum ist die Relevanz dieser Dokumente: Die Inhalte bilden jeweils die Grundlage für Entscheidungen in der Klimapolitik.

Nicht nur anhand des Zeitraums und der Medien müssen Beiträge beschränkt werden, sondern auch inhaltlich. Denn es ist weder erstrebenswert noch sinnvoll, alle Artikel der genannten Jahre in das Korpus aufzunehmen. Nicht alles, was in diesen acht Jahren in diesen zweiundzwanzig Zeitungen und zwei Fernsehsendungen Aufmerksamkeit erregte, ist relevant im Diskurs über Klimawandel. Es gilt also, Beiträge zu finden, die den Diskurs über Klimawandel thematisieren und dabei ein Gleichgewicht zwischen den Suchkriterien zu finden, so dass möglichst wenig irrelevante Artikel in das Korpus Einzug finden, ohne dass gleichzeitig diskursrelevante Artikel ausgeschlossen werden. Im Rahmen des Projektes *Changing Climate* der Universität Lancaster (Centre for Corpus Approaches to Social Science (CASS) o. J.) stellten verschiedene Universitäten Korpora zum Diskurs über Klimawandel zusammen (beispielsweise Müller o. J.). Um das Gleichgewicht zwischen noch relevanten und nicht mehr relevanten Texten zu wahren sowie eine vergleichbare Datengrundlage zu erhalten, um komparative Studien durchführen zu können, wurden die Suchwörter für die Zusammenstellung der Korpora für das Projekt mittels einer von Gabrielatos (2007) vorgeschlagenen Methode erhoben. Für das deutsche Korpus (Müller o. J.) ergaben sich die folgenden Suchwörter:

> *Klimawandel, Erderwärmung, Klimaschutz, Klimaerwärmung, Ausstoß, Energiewende, Kohlendioxid, Erwärmung, Klimakonferenz(en)*

Diese Suchwörter wurden auch für die Erstellung des Schweizer Korpus verwendet. Alle Artikel, die zwischen 2007 und 2014 in einem der gewählten Medien veröffentlicht wurden und mindestens eines der Suchwörter aufweisen, wurden in das Korpus aufgenommen.

---

2 Das Aufnehmen von mehr Fernsehformaten in das Korpus wäre wünschenswert gewesen. Zeit- und Ressourcenbegrenzungen erschwerten dieses Vorhaben. Der Miteinbezug dieser beiden Formate muss somit klar als Kompromiss gesehen werden. Die Wahl der beiden Fernsehformate wird im späteren Verlauf dieses Kapitels begründet.

4.2 Zusammensetzung

Im Korpus selbst nehmen die Tageszeitungen eine Schlüsselrolle ein. Die Anzahl Artikel, die in Tageszeitungen über den Klimawandel veröffentlicht wurden (s. Abbildung 2), fällt weit höher aus, als der blosse Vergleich mit der Auflagenstärke (s. Abbildung 3) erwarten liesse.

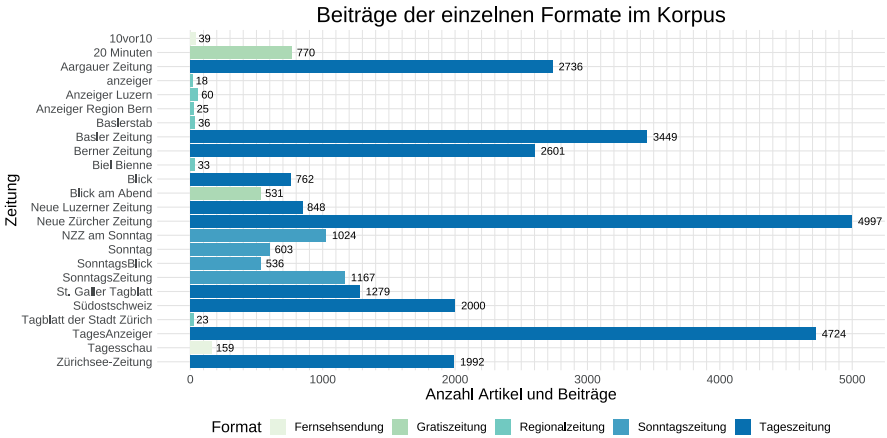

Abb. 2: Anzahl Artikel und Beiträge im Korpus in den einzelnen Zeitungen und Fernsehformaten

Abb. 3: Aufaddierte Auflagenstärke der im Korpus befindlichen Zeitungsarten von 2007 bis 2014 (gemäss Zahlen des WEMF 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014)

Tageszeitungen scheinen dementsprechend quantitativ relevant für die Vermittlung von Diskursen über Klimawandel zu sein. In Regionalzeitungen hingegen wird kaum über den Klimawandel berichtet. Zusätzlich weisen auch Zeitungen innerhalb einer Zeitungsform deutliche Unterschiede auf, so wurden in der *Neuen Zürcher Zeitung* im betrachteten Zeitraum fast sechsmal mehr Artikel über den Klimawandel publiziert als in der *Neuen Luzerner Zeitung*.

4.2.1 Schweizer Medienlandschaft

Die Schweizer Medienlandschaft ist im Umbruch und war über die letzten Jahre hinweg von drei wesentlichen Tendenzen geprägt.[3] Wenige Medienhäuser sind für einen immer grösseren Teil der Presse zuständig und die Leserzahlen traditioneller Zeitungen sind rückläufig, während sich Gratiszeitungen zunehmender Beliebtheit erfreuen (s. Publikation des Forschungszentrums Öffentlichkeit und Gesellschaft (fög 2014)).

Folgende Zeitungsarten befinden sich im Korpus:

- *Tageszeitungen* werden einmal täglich (Montag bis Freitag oder Montag bis Samstag) publiziert.
- *Sonntagszeitungen* werden einmal wöchentlich sonntags publiziert.
- *Gratiszeitungen* werden kostenlos einmal täglich im öffentlichen Raum verteilt (typischerweise von Montag bis Freitag an Bahnhöfen).
- *Regionalzeitungen* werden an Bewohner*innen einer Region kostenlos zugestellt. Teilweise handelt es sich um die amtlichen Publikationsorgane der entsprechenden Region. Die Häufigkeit variiert; oft werden Regionalzeitungen zweimal pro Woche verteilt.

Wie Abbildung 3 bereits zeigte, sind Tageszeitungen die auflagenstärksten Zeitungen im Korpus, gefolgt von Regional- und Sonntagszeitungen. Die Auflagenstärke der Gratiszeitungen ist trotz ihrer Bedeutung verhältnismässig klein, was unter anderem daran liegt, dass im Wesentlichen zwei Gratiszeitungen (*20 Minuten* und *Blick am Abend*) mehreren Tages- und Sonntagszeitungen gegenüberstehen. Zudem kann davon ausgegangen werden, dass einzelne Exemplare mehrfach gelesen werden; als sogenannte «Pendlerzeitungen» werden sie in den öffentlichen Verkehrsmitteln gelesen, liegengelassen und von nachfolgenden Passagier*innen erneut gelesen.

3 Die Medienlandschaft in der Schweiz soll an dieser Stelle nur so weit ausgeführt werden, wie dies dem/der der Schweiz unbekannten Leser*in einen allgemeinen Überblick sowie einen genaueren Blick auf die Zeitungen und Fernsehformate im Korpus ermöglicht (für einen Überblick über Besonderheiten der schweizerischen Medienlandschaft s. Studer et al. 2014).

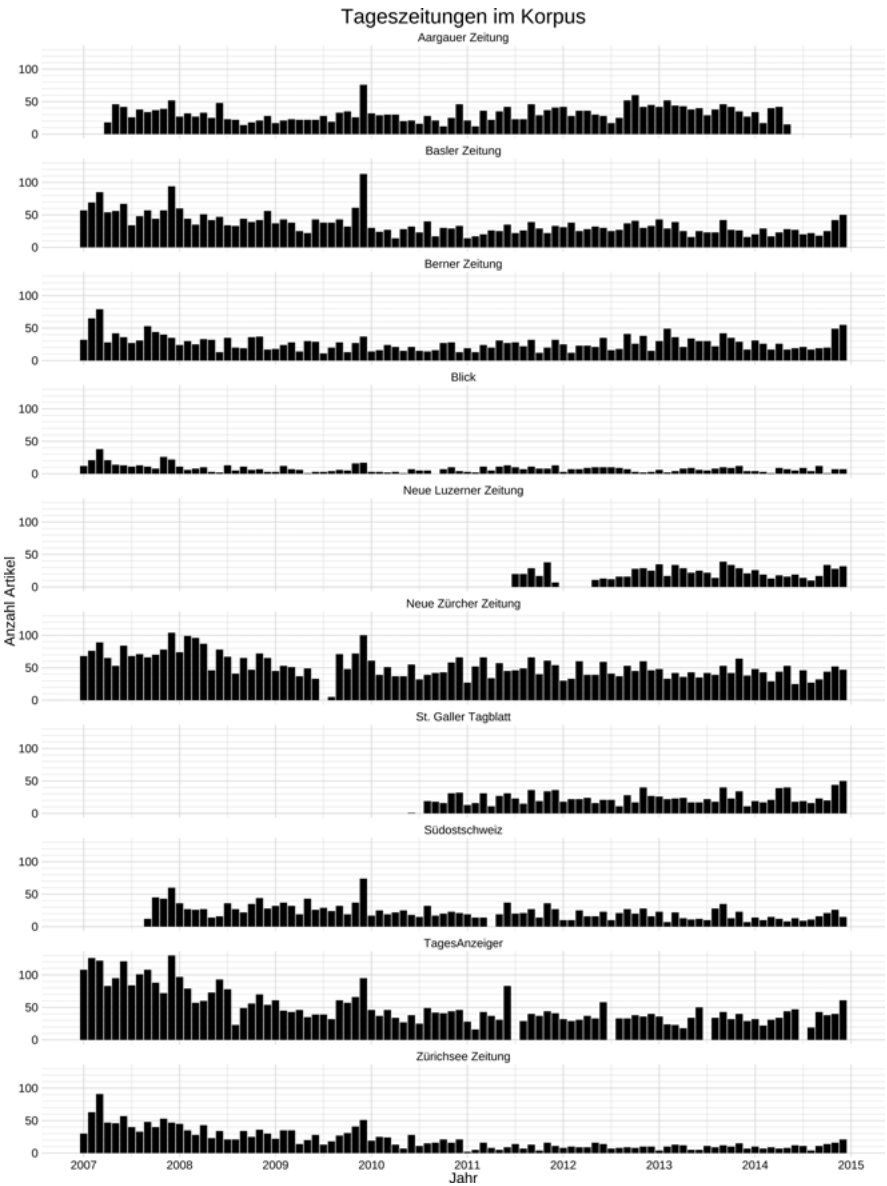

Abb. 4: Monatlich publizierte Artikel in Tageszeitungen im Korpus

4.2.1.1 Tageszeitungen

Im Korpus befinden sich zehn Tageszeitungen. Neun davon werden durch fög (2014) als Abonnementenzeitungen bezeichnet, die zehnte – der *Blick* – als Boulevardzeitung.[4] Im betrachteten Zeitraum ist die Auflagenstärke der Tageszeitungen insgesamt rückläufig (s. WEMF 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014). Dies zeigt sich auch an den im Korpus vorhandenen Artikeln (s. Abbildung 4).

Die Qualität[5] der Tageszeitungen nahm insgesamt ab (Gesamtqualität 2010: 5,1; 2013: 4,9, die Daten entstammen fög 2014: 99), diejenige der Boulevardzeitungen hingegen zu (2010: 2,6; 2013: 3,3; fög 2014: 99). Als Grund für die Qualitätsabnahme werden die derzeitigen Schwierigkeiten im Medienbereich genannt: Finanzielle Umstände führen dazu, dass Einordnungsleistung und Beitragsrelevanz tendenziell abnehmen. Die qualitative Verbesserung der Boulevardpresse resultiert hingegen aus einer Versachlichung (s. fög 2014: S. 98). Ein auch im Korpus beobachtetes Phänomen ist, dass Artikel oft auf Agenturmeldungen basieren (s. fög 2014: 37). In der Konsequenz lassen sich Artikel in gleicher oder ähnlicher Form in verschiedenen Zeitungen finden, wodurch die Vielfalt insgesamt abnimmt.

4.2.1.2 Sonntagszeitungen

Dem gegenüber stehen vier Sonntagszeitungen, wobei eine – der *SonntagsBlick* – der Boulevardkategorie zugeordnet wird. Die Auflagenstärke von Sonntagszeitungen ist insgesamt höher und über die Jahre hinweg stabiler als diejenige der Tageszeitungen (s. WEMF 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014). Im Korpus schwankt die Berichterstattung hingegen, wie die monatliche Verteilung in Abbildung 5 verdeutlicht.

---

4 Die Unterscheidung von Abonnements-, Boulevard und Sonntagspresse, die in fög (2014) verwendet wird, wird im Rahmen dieser Arbeit nicht übernommen, da zahlreiche Zeitungen mehreren dieser Kategorien angehören. Zudem erscheint mir die Unterscheidung von Abonnements-, Boulevard- und Sonntagszeitungen inkonsistent; die Distinguierung erfolgt aufgrund der Art der Bezahlung, des Inhalts sowie der Anzahl wöchentlicher Ausgaben.

5 «Dieses Verständnis [von Qualität] geht ursprünglich auf den Aufklärungsliberalismus zurück und manifestiert sich seither in den Ansprüchen auf Universalität, Relevanz, Ausgewogenheit und im Objektivitätsstreben beim öffentlichen Räsonieren als Voraussetzung für eine funktionierende Demokratie. Diese Ansprüche finden sich wieder in den modernen Qualitätsnormen der ‹Vielfalt›, der ‹Relevanz›, der mit ihr verbundenen ‹Aktualität› und in den wesentlichen Anforderungen an die ‹Professionalität› journalistischen Arbeitens (u. a. Sachlichkeit, Eigenleistung, Quellentransparenz).» (fög 2014: 28)

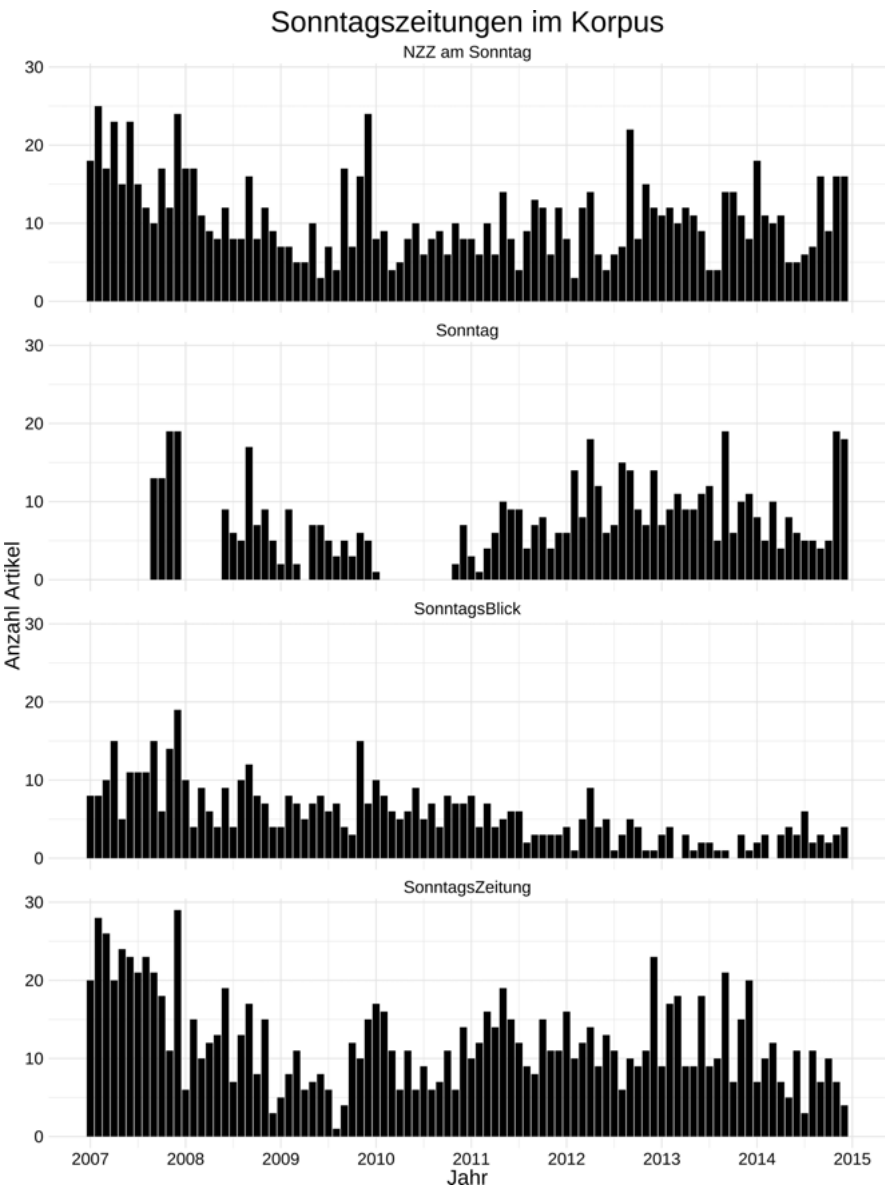

Abb. 5: Monatlich publizierte Artikel in Sonntagszeitungen im Korpus

Nicht nur die Auflagenstärke, sondern auch die Qualität der Sonntagszeitungen ist verhältnismässig stabil (2010: 4,8 sowie 2013: 4,7; fög 2014: 99), wobei sich

insbesondere der *SonntagsBlick* in diesen drei Jahren stark verbesserte (plus 0,8 Punkte; fög 2014: 99–100). Diese Unterschiede zu den Tageszeitungen lassen sich unter anderem damit erklären, dass es bei Sonntagszeitungen oft noch typische Redaktionen gibt, welche die Artikel der Kernressorts schreiben. Gleichzeitig besteht aber die Tendenz, dass in besagten Redaktionen stets mehr Zeit für *Softnews*[6] aufgewendet wird (s. fög 2014: 37).

4.2.1.3 Gratiszeitungen

Im Korpus befinden sich zwei kostenlose Zeitungen: *20 Minuten* und *Blick am Abend* (s. Abbildung 6).

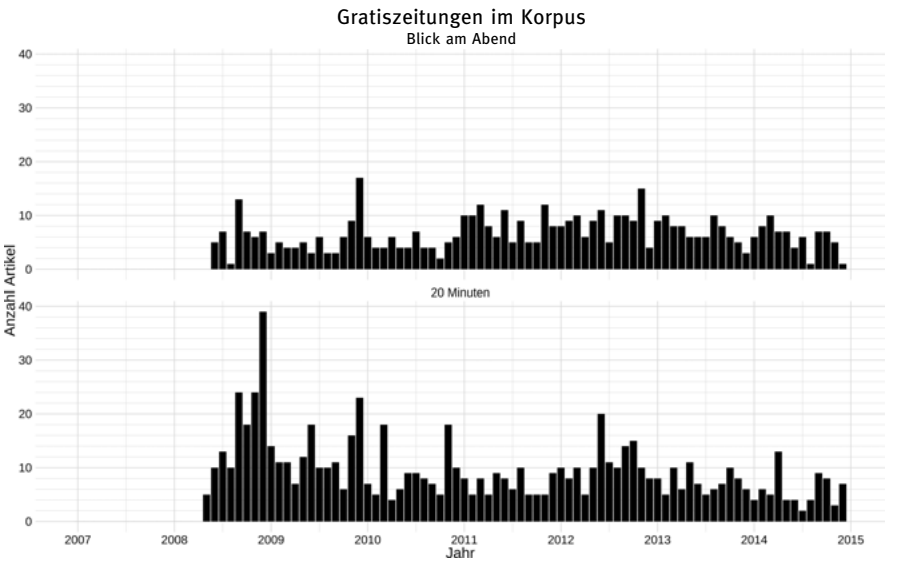

Abb. 6: Monatlich publizierte Artikel in Gratiszeitungen im Korpus

Diese Gratiszeitungen sind mindestens in der Schweiz ein relativ neues Phänomen; so gibt es *20 Minuten* seit 1999, *Blick am Abend* erst seit 2008.[7] Weitere kostenlose Zeitungen wurden publiziert, die Verbreitung wurde allerdings wieder eingestellt. Die Präsenz von Gratiszeitungen unterscheidet die Schweiz auch vom europäischen Umland, da sie in der Schweiz «zu den auflagen- und reich-

6 «Weiche Nachrichten (‹soft news›) sind solche, deren Nachrichtenwert bei eher fehlender objektiver Bedeutung vor allem durch die Neugier und Sensationslust des Publikums bestimmt wird.» (Mast 2018: 347) Im Gegensatz dazu stehen *Hardnews*.

7 Die Produktion von *Blick am Abend* wurde am 21. Dezember 2018 eingestellt (Ringier 2018).

weitenstärksten Titeln der Presse» (fög 2014: 30) gehören. Durch die enorme Reichweite erreichen diese beiden Zeitungsformate grosse Teile der Öffentlichkeit:

> Betrachtet man die Bildungs- und Einkommensverteilung, so ist diese bei den Gratiszeitungen praktisch identisch mit dem Durchschnitt der (mediennutzenden) Bevölkerung. Dies ist Ausdruck der ausgesprochen hohen Reichweite und gesellschaftlichen Verbreitung, die die Gratiszeitungen mittlerweile über alle Bildungs- und Einkommensschichten hinweg erreicht haben.
>
> (fög 2014: 38–39)

Gratiszeitungen sind als qualitativ schlecht (2010: 3.1 und 2013: 3.1, s. fög 2014: 39) einzuordnen, unter anderem, weil «die Hardnews in Gratiszeitungen [...] überdurchschnittlich oft lediglich umgeschriebene Agenturmeldungen» sind (fög 2014: 37).

4.2.1.4 Regionalzeitungen

Unter dem Terminus *Regionalzeitungen* werden an dieser Stelle verschiedene Zeitungsformate subsumiert. Es handelt sich dabei einerseits um lokale Blätter, die kostenlos an die Anwohner*innen einer bestimmten Region gesendet werden, und andererseits um amtliche Publikationsorgane. Die Auflagenstärke der Regionalzeitungen hängt stark von der Anzahl an Einwohner*innen der Regionen ab. Regionalzeitungen erscheinen oft wöchentlich, so der *Anzeiger Luzern*, *anzeiger – Das Ostschweizer Wochenmagazin*, *Biel Bienne*[8] sowie das amtliche Publikationsorgan der Stadt Zürich *Tagblatt der Stadt Zürich*. Das amtliche Publikationsorgan der Stadt und Agglomeration von Bern – *Anzeiger Region Bern* – erscheint hingegen zweimal wöchentlich. Der mittlerweile eingestellte *Baslerstab* konnte im öffentlichen Raum Zeitungsboxen entnommen werden. Zu Beginn des Zeitraums wurde er noch bis zu fünfmal pro Woche ausgetragen, gegen Ende nur noch wöchentlich. Die *BaZ kompakt* ist der unmittelbare Nachfolger des *Baslerstabs*. Aus Gründen der Zugänglichkeit befindet sie sich aber nicht im Korpus. Im betrachteten Zeitraum fiel die Berichterstattung über den Klimawandel in den Regionalzeitungen eher gering und punktuell aus (s. Abbildung 7).

8 Hierbei handelt es sich um eine Gratiszeitung, die im Kanton Biel vertrieben wird. Die Zeitung ist zweisprachig – Deutsch und Französisch – aufgebaut. Im Korpus befinden sich nur die deutschsprachigen Artikel.

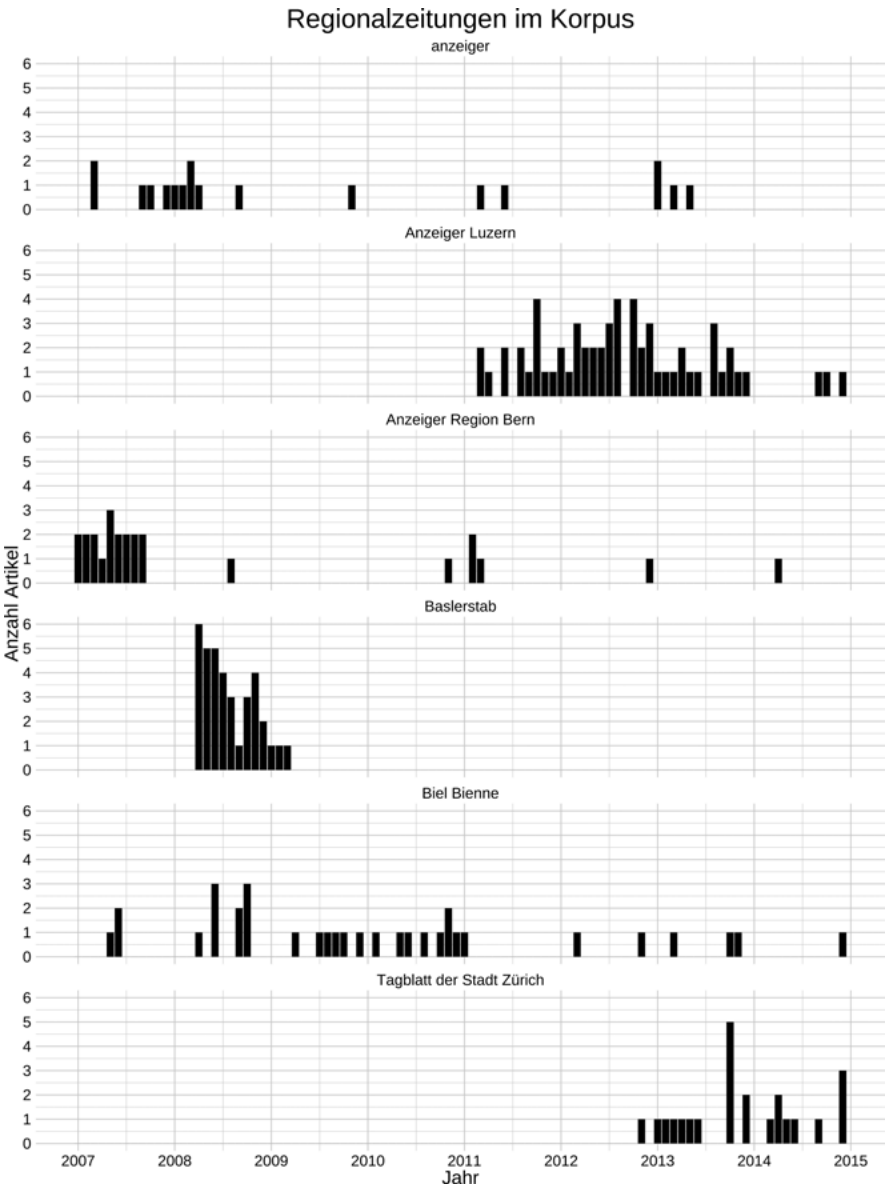

Abb. 7: Monatlich publizierte Artikel in Regionalzeitungen im Korpus

4.2.1.5 Fernsehformate

Da das Korpus den öffentlichen Mediendiskurs repräsentieren soll, wurden zwei Fernsehnachrichtenformate gewählt: *10vor10* und *Tagesschau*. Die *Tagesschau* informiert viermal täglich über «Themen aus Politik, Wirtschaft, Kultur, Sport, Gesellschaft und Wissenschaft. Sie gibt einen Überblick über die wichtigsten Ereignisse des Tages» (Schweizerische Radio- und Fernsehgesellschaft (SRG) o. J.a) und ist eine der beliebtesten und meistgesehenen Sendungen der Schweiz.[9] Bei *10vor10* handelt es sich um die zweite Nachrichtensendung der öffentlich-rechtlichen Fernsehsender. Diese Sendung soll einmal täglich vertiefte Hintergrundinformationen bieten (SRG o. J.b). Das Anwenden des Zeitraums sowie der Suchwörter führt dazu, dass sich 39 Beiträge aus *10vor10* und 159 Artikel aus der *Tagesschau* im Korpus befinden (s. Abbildung 8).[10]

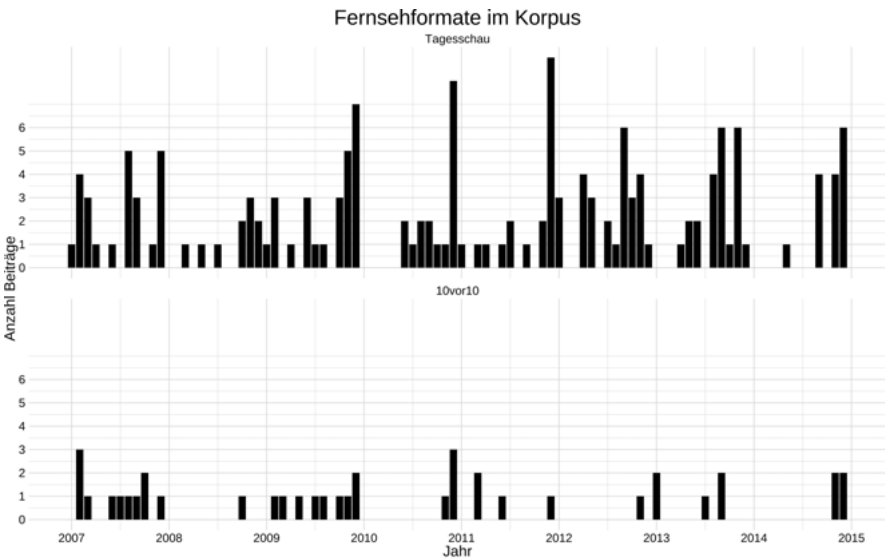

Abb. 8: Monatlich ausgestrahlte Fernsehsendungen im Korpus

9 Eine Begebenheit, welche Rebetez (2006) schildert, zeigt die Bedeutung der *Tagesschau* auf. Eine im deutschsprachigen Raum angewandte Methode zur Messung der durchschnittlichen Tagestemperatur war es, Personen dreimal täglich zu bestimmten Zeiten die Temperaturen an einer bestimmten Stelle messen zu lassen (Rebetez 2006: 35–36): «Bis 1970 wurde in der Schweiz um 7.30 Uhr, 13.30 Uhr und 21.30 Uhr gemessen. Ab 1971 wurden die Messzeiten dann auf 7.00 Uhr, 13.00 Uhr und 19.00 Uhr verlegt. […] Da es ohnehin schwierig war, Freiwillige für diese schlecht bezahlte, einschränkende und nicht unbedingt dankbare Aufgabe zu finden, beschloss man, den dritten Messzeitpunkt auf den frühen Abend zu verlegen, vor den Beginn der Tagesschau.»

10 Die Beiträge entstammen der Plattform *PLAYSRF* (SRG o. J.e), auf der die meisten Sendungen der öffentlich-rechtlichen Schweizer Kanäle kostenlos zugänglich sind.

Mit Hilfe von Entwürfen, die mir die beiden Redaktionen dankenswerterweise zur Verfügung stellten, wurde dann eine Transkription der Sendungen vorgenommen. Schweizerdeutsche sowie französische Passagen wurden aufgrund der Zugänglichkeit in den schriftsprachlichen Standard übersetzt. Die Transkription orientiert sich stark am geschriebenen Standard, da dies weniger zeitintensiv und auch korpuslinguistisch besser greifbar ist.

4.3 Aufbereitung

Das Sammeln von Artikeln reicht keineswegs für korpusanalytische Zugänge aus, da diese für Programme ohne Weiterverarbeitung noch nicht lesbar sind und Kontextinformationen fehlen. Deshalb müssen Artikel (semi-)automatisch sortiert und annotiert werden. Diese Vorgänge machen die Daten eines Korpus einerseits zugänglich, limitieren sie andererseits aber auch, denn hinter jeder Annotation stecken Entscheidungen, die auf theoretischen und methodologischen Annahmen fussen und somit eine gewisse Perspektive auf das Korpus nach sich ziehen, während dadurch andere in den Hintergrund rücken.

Die Beiträge aus dem Korpus stammen aus unterschiedlichen Datenbanken: *LexisNexis* (o. J.), *Factiva* (o. J.), *SwissDox* (o. J.), redaktionsinterne Archive sowie *SRFPlay* (SRG o. J.e) sind vertreten. Jede einzelne hat eigene Darstellungsformate und Metadaten; selbst innerhalb der einzelnen Datenbanken sind die Darstellungen der Artikel und der Metadaten teilweise heterogen aufgrund der Länge des betrachteten Zeitraums. Dies erschwert das Unterfangen, Beiträge in ein mit der *Corpus Workbench* (CWB) (Evert & Hardie o. J.) kompatibles Format zu überführen[11] und möglichst viele Kontextinformationen zugänglich zu machen, beträchtlich, da eine manuelle Annotation von Metadaten aufgrund der Beitragsmenge weder erstrebenswert noch realistisch ist.

Im Folgenden soll ein Überblick über diejenigen Informationen gegeben werden, die sich automatisiert aus den digitalen Daten erheben liessen und somit auch Eingang in das Korpus fanden. In den Datenbanken liegen die Artikel in vereinfachter Form vor, die je nach Datenbank mehr oder minder stark von den tatsächlichen Zeitungsartikeln abweichen. Spalten, (oft auch) Bilder, Schriftarten und -grösse, Platzierung sowie Layout etc. gehen in der Regel verloren. Für alle Dokumente sind Datum, Zeitung, der Fliesstext sowie der Titel zugänglich, seltener auch das Ressort, die Seitenangabe des Artikels, der Lead

11 Zu den für die *Corpus Workbench* benötigten Formaten s. Evert & CWB Development Team (2016).

und der Autor. Bildunterschriften, Zwischentitel oder Zitate lassen sich hingegen nur selten als solche identifizieren.

Diese Informationen werden in ein für das Programm lesbares Format übertragen. Die Struktur eines Zeitungsartikels im Korpus besteht aus den in Abbildung 9 dargestellten Elementen.

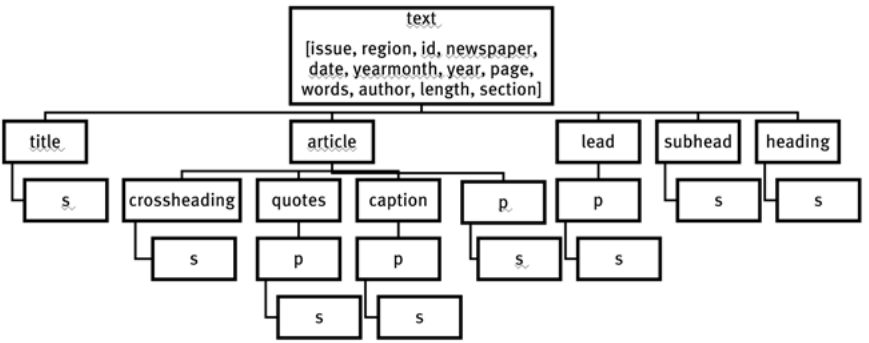

Abb. 9: XML-Syntax der Zeitschriftenartikel. Mit Wellenlinie sind diejenigen Informationen versehen, die alle Artikel aufweisen. *p* sind Absätze, *s* einzelne Sätze.

Die Codierung der Fernsehbeiträge weicht von diesem Schema ab, da Fernsehbeiträge anders aufgebaut sind und teilweise auch andere Metadaten aufweisen. Nebst Moderatorenbeiträgen (*moderator*) können auch Beiträge mit einer Stimme aus dem Off (*input*) und Sprecherbeiträge (*quote*) unterschieden werden (s. Abbildung 10).

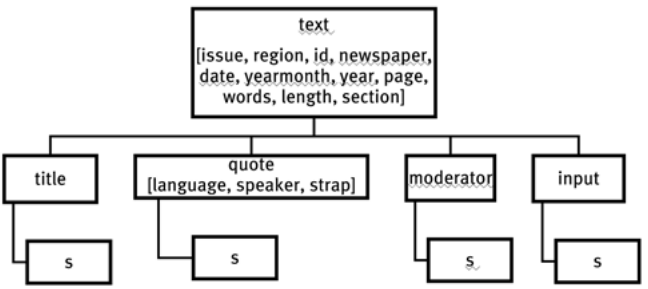

Abb. 10: XML-Syntax der Fernsehsendungen. Mit Wellenlinien sind diejenigen Informationen versehen, die alle Sendungen aufweisen. Da mündliche Beiträge keine Absätze aufweisen, werden nur Sätze (*s*) markiert.

Wie bereits erwähnt, unterschieden sich die einzelnen Datenbanken bezüglich der Zugänglichkeit und Vollständigkeit von Metadaten stark. Deshalb sind auch nicht für alle Artikel sämtliche Metadaten verfügbar, was die Tabelle 1 verdeutlicht.

Tab. 1: Prozentuales Vorkommen der Metadaten im Korpus

Metadatum	Anzahl Artikel mit Metadatum
Zeitung/Fernsehformat	100 %
Datum	100 %
Seite	82.1 %
Ressort[12]	78.6 %
Anzahl Wörter	77.5 %
Autor*in	49.3 %
Region	4.3 %
Ausgabe	0.3 %

Dies hat zur Folge, dass auf Metadaten beruhende Analysen ausser bei Zeitung und Datum stets nur Teilanalysen darstellen.

4.3.1 Annotationen im Korpus

Die Vorbereitung von Korpora endet selten bei den eben vorgestellten Metadaten, denn weitere Annotationen können entweder zusätzliche Zugriffsmöglichkeiten auf das Korpus bieten (wie beispielsweise mittels Lemmatisierung) oder aber weitere Aspekte der Kontexte greifbar machen (wie beispielsweise Themen):

> Annotationen stellen normalerweise *Generalisierungen* dar, z. B. wird von einzelnen Wortinstanzen auf allgemeine Klassen abstrahiert wie auf die Wortart *Adjektiv*. Generalisierungen wiederum helfen, wiederkehrende *Muster* in der Vielfalt der Formen zu erkennen. Bei der Suche dienen Annotationen als *Anker*, über den die annotierten Phänomene *effizient* und *nachhaltig* wieder auffindbar sind.
>
> (Zinsmeister 2015: 86)

12 Ressorts sind zusätzlich bereits in den konsultierten Datenbanken teilweise falsch verschlagwortet. Deshalb ist eine Verortung innerhalb der Ressorts nur in Einzelfällen möglich.

Mit Blick auf die Menge an Artikeln sind insbesondere Methoden der automatischen Annotation interessant, da eine händische Annotation über alle Artikel hinweg im Rahmen der Arbeit nicht möglich wäre. Solche automatischen Annotationen können aber gerade mit Blick auf eine Analyse aus der Perspektive des Diskurshistorischen Ansatzes problematisch sein: Automatische Annotationen sind insbesondere für das zeitgenössische Englisch verfügbar. Bewegt man sich diachron oder mit Blick auf die Sprache von diesem Standard weg, so schwinden auch die Möglichkeiten für automatische Annotation. Weiter sollte berücksichtigt werden, dass solche Annotationen immer auf theoretischen Konzepten beruhen, die sich nicht zwingend mit denjenigen decken, die in den Analysen bevorzugt werden. Um diesen beiden Punkten zu begegnen, habe ich teilweise auch individuelle Annotationen angewandt. Dies bedeutet, Kategorien wurden aufgrund von theoretischen Konzepten etabliert, für eine automatische Annotation generalisiert und dann auf das Korpus angewandt. Dies führt selbstredend nicht zu solch komplexen Annotationen wie durch automatische Annotationen. Hierfür wären allenfalls *Machine Learning*-Ansätze hilfreich, die aber im Rahmen dieser Studie nicht thematisiert werden.

Die Annotation des Korpus werde ich aus den genannten Gründen im Rahmen des nachfolgenden Kapitels 5 besprechen, in dem auch die sektorale Argumentationstheorie erörtert wird, um den Anschluss der Annotation an die Entwicklung einer solchen Argumentationstheorie gewährleisten zu können. Denn die Wahl der Annotationen hängt stark von der Konzeptualisierung der sektoralen Argumentationstheorie, die der Beantwortung der Fragestellung dienen soll, ab; dies gilt insbesondere für manuell etablierte Kategorien. Dementsprechend scheint es mir, auch wenn es eher unüblich ist, sinnvoller, die Annotationen im nachfolgenden Kapitel an den entsprechenden Stellen zu verorten, damit sie nicht theorielos im Raum stehen. Es handelt sich dabei um folgende Annotationen:

- Metadaten der Zeitungsartikel und Fernsehbeiträge (Annotation von Texten):
- Sequenzierung, Lemmatisierung, Wortarten mittels *TreeTagger* (Schmid o. J.) und *Stuttgart-Tübingen Tagset* (Schiller et al. 1999): Kapitel 4.3.1.1
- 90 mittels *Mallet* (McCallum 2002) erhobene Themen (Annotation von Wörtern und Texten): Kapitel 5.1.1
- Modalverben anhand der von Helbig & Helbig (1990) vorgeschlagenen Kategorien sowie Realisierungsmöglichkeiten für Negation (Zifonun, Hoffmann & Strecker 1997: 147): Kapitel 5.2.3
- Eigennamen mittels *Stanford NER* (Stanford NLP Group o. J.): Kapitel 5.3.4
- Statistische Erhebung von *Keywords* und *Co-Keywords* im Korpus: Kapitel 5.4.1.2.1

- Pragma-dialektische Indikatorenfür unterschiedlichen Argumentationsphasen und -schemata anhand von van Eemeren, Houtlosser & Snoeck Henkemans (2005) sowie van Eemeren, Houtlosser & Snoeck Henkemans (2007): Kapitel 5.4.2
- Beschreibende und meinungsbasierte Texte anhand der annotierten Ressorts: Kapitel 6.3

Die Sequenzierung, Lemmatisierung und Annotation von Wortarten wird sogleich thematisiert, handelt es sich doch um ein Verfahren, das überhaupt erst einen sinnvollen Zugriff auf das Korpus ermöglicht und somit in der Regel unabhängig vom gewählten Ansatz angewandt wird.

4.3.1.1 Sequenzierung, Lemmatisierung und Annotation von Wortarten

Ein Standardverfahren, das zur Aufbereitung von Korpora verwendet wird, ist die Wortartenannotation. Häufig findet gleichzeitig eine Sequenzierung von Sätzen sowie eine Lemmatisierung statt, da beides in der Regel notwendig für die Annotation von Wortarten ist. Allerdings können diese beiden Schritte bereits problematisch sein, so kann beispielsweise die Lemmatisierung von Verben mit trennbaren Präfixen Schwierigkeiten bereiten und bei der Sequenzierung muss die nicht ganz triviale Frage nach Satzgrenzen beantwortet werden. Im vorliegenden Korpus wurden die Satzgrenzen gemäss dem Stuttgart-Tübingen Tagset (im Folgenden STTS; Schiller et al. 1999: 73) festgelegt.[13] Für die Annotation selbst existieren regel- und statistikbasierte Systeme (s. Zinsmeister 2015), zu letzteren gehört auch das im deutschsprachigen Raum häufig und auch hier verwendete Programm *TreeTagger* (Schmid o. J.) mit dem STTS (Schiller et al. 1999). Es basiert auf einer traditionellen Einteilung von Wortarten. Ein Tagger, der sich an der systemisch-funktionalen Grammatik (s. Halliday & Matthiessen 2014) orientierte, wäre aus Sicht des Diskurshistorischen Ansatzes wünschenswert und zeigt deutlich, dass korpuslinguistische Methoden trotz ihrer Möglichkeiten durchaus auch einschränkend sein können.

Die automatische Annotation von Wortarten ist anfällig für Fehler (offensichtlich bei Homonymen oder unbekannten Wörtern). Zwar ist die korrekte Zuordnung von Wortarten durch den *TreeTagger* insgesamt recht hoch,[14] allerdings treten regelmässig anwendungstypische, systematische Fehler auf. An dieser Stelle sollen die Funktionsweisen des *TreeTaggers* kurz erläutert und

13 Demnach markieren Punkte, Ausrufe- und Fragezeichen, Semikola sowie Doppelpunkte Satzgrenzen.

14 Zinsmeister (2015: 96) spricht von einer Akkuratheit von 95,82 % pro Token für den *TreeTagger*.

mögliche Fehlerquellen identifiziert werden, denn «[d]ie Fehler, die diese Tools produzieren, haben bei genauerer Betrachtung gemeinsam, dass sie systematisch auftreten und in gewisser Weise erklärbar und auch vorhersehbar sind» (Zinsmeister 2015: 89). Das Wissen um solche systematischen Fehler ist somit für Analysen äusserst wichtig, da sie bei entsprechenden Suchanfragen beachtet und im Idealfall umgangen werden müssen.

TreeTagger arbeitet einerseits mit einer Wortliste, in der Lexeme einer entsprechenden Wortarten-Wahrscheinlichkeit zugeordnet sind, anderseits mit Sequenzen von bis zu drei Tags. Dieser sequenzbasierte Ansatz soll anhand eines Beispiels aus Zinsmeister (2015: 91–92) erläutert werden: *Erhalten* tritt am häufigsten als Infinitiv eines Vollverbs (in der Terminologie des STTS gesprochen als VVINF) und am zweithäufigsten als Partizip eines Vollverbes (VVPP) auf. Betrachtet man 3-Gramme mit *erhalten*, so treten solche der Abfolge Nomen–*zu* (vor Infinitiv)–Infinitiv eines Vollverbs (NN–PTKZU–VVINF, beispielsweise «den Friedhof zu erhalten») am häufigsten auf, gefolgt von der Kombination Artikel–Nomen–Partizip Perfekt eines Vollverbs (ART–NN–VVPP, beispielsweise «das Landeskriminalamt Berlin habe einen Hinweis erhalten»). Durch die Kombination von Wortlisten und diesem sequenzbasierten Ansatz kann die Akkuratheit insgesamt erhöht und die Annotation unbekannter Wörter ebenfalls ermöglicht werden. Die Berechnung sequenzbasierter Wahrscheinlichkeit ist beim *TreeTagger* mit Blick auf zwei Punkte eingeschränkt. Erstens führt die Reduzierung auf drei Lexeme dazu, dass eindeutige Konstruktionen wie «Der Ständerat will Offroader in der Schweiz nicht verbieten.» (10vor10 07.03.2011) nicht berücksichtigt werden. Interessiert man sich für die Wortart von *verbieten*, so ist der Blick auf *will* äusserst dienlich. Diesen Schluss kann *TreeTagger* allerdings nicht ziehen, da er für die wahrscheinlichkeitsbasierte Verteilung nur die drei Tokens *Schweiz nicht verbieten* berücksichtigt. Zweitens wird immer nur die Wahrscheinlichkeit des letzten Ausdrucks in einem von links nach rechts gelesenen 3-Gramm berechnet. In dem Beispiel «Die EU will anscheinend nichts mehr wissen von ihrem CO_2-Reduktionsziel von 30 %» (10vor10 16.12.2009; Hervorhebungen N. K.) werden *von ihrem* bei der Annotation von *wissen* nicht beachtet. Weitere Schwierigkeiten und Besonderheiten müssen bei der Verwendung von *TreeTagger* beachtet werden (Zinsmeister 2015: 105–106):

- Nomen können nicht klar von Eigennamen,
- adverbial genutzte Adjektive nur schwer von abgeleiteten Adverbien,
- und prädikativ genutzte Adjektive nicht immer von verbalen Partizipien abgegrenzt werden.
- Zusätzlich werden *sein* und *haben* immer als Hilfsverben klassifiziert.

Während der Analyse hat sich zudem gezeigt, dass die Annotation (und Lemmatisierung) trennbarer Verben ebenfalls problematisch ist.