

bt_reactions
Zwischenreaktionen im 17.-19. Bundestag:
Datenhandbuch V1.0

Lukas Höttges & Jakob Tures

22. November 2023

Inhaltsverzeichnis

1 Datensatz	3
1.1 Inhalt	3
1.2 Erstellung	3
1.3 Beteiligte	4
2 Variablen	4
2.1 Beobachtungsbezogen	5
2.1.1 ID	5
2.1.2 Periode	5
2.1.3 Sitzung	5
2.1.4 Datum	6
2.2 Zwischenreaktion	6
2.2.1 Typ	6
2.2.2 Zwischenruf	6
2.2.3 Reaktion	7
2.3 Codierung	7
2.3.1 Sampled	7
2.3.2 Negativ	8
2.3.3 Sexism	8
2.3.4 Incivility	9
2.3.5 Bezug	10
2.4 ZwischenruferIn	10
2.4.1 Name	10
2.4.2 Fraktion	11
2.4.3 Regierungsfunktion	12
2.4.4 Fraktionsfunktion	13
2.4.5 Gender	13
2.4.6 Alter	14
2.4.7 Akad_Titel	14
2.5 Fraktionsbezogen	15
2.5.1 Regierungsfraktion	15
2.5.2 Fraktionsgroesse	16
2.6 Reaktion auf	16
2.6.1 Reacted_Name	16
2.6.2 Reacted_Fraktion	17
2.6.3 Reacted_Regierungsfraktion	18
2.6.4 Reacted_Regierungsfunktion	18
2.6.5 Reacted_Gender	19
2.7 Gewichte	20
2.7.1 PW	20
2.7.2 IPW	20
3 Anwendung der Gewichte mit survey	20

1 Datensatz

1.1 Inhalt

Der Datensatz `bt_reactions_1_0.RData` umfasst alle 824,996 Zwischenreaktionen¹ aller Bundestagssitzungen der 17.–19. Legislaturperioden des Deutschen Bundestages sowie die zugehörigen Metadaten, Daten zu der reagierenden Person bzw. Fraktion sowie die Daten zu der unterbrochenen Person bzw. Fraktion. Zwischenreaktionen sind dabei neben Zwischenrufen auch die *weiteren Reaktionen: Beifall, Heiterkeit, Lachen und Widerspruch*². Der Datensatz umfasst insgesamt 253,010 Zwischenrufe und 571,986 weitere Reaktionen.

Für eine geschichtete Zufallsstichprobe von 17,999 Zwischenrufen liegen außerdem manuelle Codierungen des Textinhalts in den Kategorien *Negativität, Inzivilität, Sexismus* und *Bezug* (Inhaltlich oder Persönlich) vor³.

Der Datensatz ermöglicht somit im Allgemeinen die Analyse des Zwischenreaktionsverhaltens der Fraktionen des Deutschen Bundestages auf Basis aller Zwischenreaktionen der 17-19. Legislaturperiode. Im Speziellen sind inferenzstatistische Analysen zu den codierten inhaltlichen Kategorien der Zwischenrufe realisierbar. Analysen sind dabei für einzelne Legislaturperioden, vergleichend zwischen Legislaturperioden, über den gesamten abgebildeten Zeitraum sowie für nutzerdefinierte Zeiträume umsetzbar.

1.2 Erstellung

Der Deutsche Bundestag veröffentlicht alle Protokolle der Plenarsitzungen unter <https://www.bundestag.de/protokolle> in den Formaten `.pdf`, `.txt` und `.xml`. Leider ist die Maschinenlesbarkeit der angebotenen Daten noch als problematisch einzuschätzen. Die XML Daten sind erst ab der 19. Legislaturperiode ausreichend vorstrukturiert, um direkt einlesbar zu sein. Für früherer Legislaturperioden gleichen die XML Daten den `txt` Dateien in so fern, als dass jeweils das gesamte Protokoll einer Sitzung als weitestgehend unstrukturierter Fließtext vorliegt. Problematisch ist dabei vor allem, dass Kopf- und Fußzeilen sowie nicht zum eigentlichen Text gehörende Annotationen⁴ der als PDF veröffentlichten Protokollseiten Teil des Fließtextes sind.

Aufgrund der problematischen Textqualität dieser Formate, fiel die Entscheidung die PDF Versionen der Protokolle als Ausgangsmaterial zu verwenden. Dazu wurden zunächst alle Protokolle des 17.-19. Bundestages mit Web Scraping Methoden⁵ von der Website des Bundestages heruntergeladen und archiviert. Diese Dateien wurden unter Linux in einem Bash-Script⁶ zunächst in Bilddaten umgewandelt. In diesem Arbeitsschritt konnte der Bildausschnitt so festgelegt werden, dass alle Fuß- und Kopfzeilen sowie Annotationen bereits entfernt waren⁷. Das Script extrahierte im Anschluss den Text der Protokolle mit *Optical Character Recognition* (OCR) aus diesen Bilddaten und speicherte diesen als `txt`.

Die Textdateien wurden in R eingelesen und unter extensiver Nutzung von *Regular Expressions* in einen Datensatz überführt. Dabei wurde der Fließtext in einzelne Reden und die dazugehörigen Zwischeneaktionen zerlegt sowie Name, Fraktion, Titel und parlamentarische Funktion der RednerInnen bzw. reagierenden MdB⁸ ausgelesen. Dieser iterative Prozess wurde von ausführlichen Qualitätskontrollen begleitet. Durch den OCR-Prozess sowie durch Formatierungsfehler in den original Protokollen entstandene Probleme wurden teilweise automatisiert, teilweise manuell korrigiert.

Diesem Datensatz wurden personen- und fraktionsbezogene Informationen aus den Stamm-

¹Aus technischen Gründen konnte Sitzung 250 der 17. Legislaturperiode nicht in den Datensatz aufgenommen werden. Dass diese Sitzung somit Missing ist, sollte auf substantielle Analyseergebnisse auf Basis dieses Datensatzes nur minimalen oder keinen Einfluss haben.

²Erläuterungen zu den Arten von *weiteren Reaktionen* in Abschnitt 2.2.3.

³Siehe Abschnitt 2.3

⁴Dies sind 4 mit A-D markierte Quadranten der Seite, sowie der Name der RednerIn oberhalb des eigentlichen Textes.

⁵Genutzt wurde das R-Package `RSeelenium`.

⁶Durchgeführt wurde die Umwandlung in Textdaten unter Ubuntu 20.04. Die PDF wurden mit ImageMagick in TIFF Dateien umgewandelt. Aus diesen wurde der eigentliche Text mit der OCR-Engine Tesseract extrahiert.

⁷Da einzelne Dateien leicht von dem Standard-Format der Protokolle abweichen, musste hier mehrfach manuell nachjustiert werden.

⁸Mitglied des Bundestages

daten der MdB⁹ zugespielt. Nach erneuten Fehlerkorrekturen auf Basis von Unstimmigkeiten zwischen den ausgelesenen Daten und den Stammdaten der MdB wurde ermittelt auf welche Rede bzw. welches MdB mit einer Zwischenreaktion reagiert wurde bevor der Datensatz auf die Zwischenreaktionen reduziert wurde.

Zur Ermittlung der inhaltlichen Ausrichtung der Zwischenrufe, wurde im Rahmen eines Lehrforschungsprojekts an der Universität Potsdam vom SoSe 2022 bis WiSe 2022/23 ein Codierschema mit den Kategorien *Negativität*, *Inzivilität*, *Sexismus* und *Bezug* (inhaltlich oder persönlich) entwickelt. Das Schema wurde iterativ in drei Pretests weiter optimiert. Als Qualitätsmaßstab wurde neben dem Primat der inhaltlichen Sinnhaftigkeit die Intercodierreliabilität, gemessen mit Fleiss Kappa, herangezogen. Das finale Schema wurde von acht CodiererInnen¹⁰ auf eine geschichtete Zufallsstichprobe von 18.000 Zwischenrufen angewandt. Die Stichprobe wurde so geschichtet, dass pro Legislaturperiode 6.000 Zwischenrufe und innerhalb einer Legislaturperiode die gleiche Anzahl von Zwischenrufen pro Fraktion zufällig gezogen wurden¹¹. Die Grundgesamtheit der Stichprobe waren alle Zwischenrufe der 17.–19. Legislaturperiode für die der Text des Zwischenrufs vorlag und die rufende Fraktion (ohne fraktionslose MdB) bekannt war¹². Die Codierung konnte für eine Stichprobe von 17.999 Zwischenrufen realisiert werden, da ein Element der Stichprobe erst in der Codierung als False Positive erkennbar wurde, also kein Zwischenruf sondern ein fehlerhaftes Artefakt des OCR-Prozesses war¹³. Zur Kontrolle der Codierungsqualität wurden 100 Zwischenrufe von allen acht CodiererInnen bearbeitet. Die somit ermittelbare Intercodiererreliabilität wird jeweils in Abschnitt 2.3 berichtet. Während der Codierung waren den CodiererInnen keine Informationen zu den ZwischenruferInnen sowie den unterbrochenen RednerInnen (bspw. die Fraktionszugehörigkeit) bekannt, um eine Beeinflussung der Codierung zu vermeiden. Um Analysen auf Basis der geschichteten Zufallsstichprobe verzerrungsfrei zu ermöglichen, wurden Gewichte berechnet und dem Datensatz hinzugefügt (Siehe Abschnitt 2.7).

Der R-Datensatz ist als `Tibble` angelegt. Das Einlesen als Data Frame ist möglich, das Laden des R-Package `tidyverse` vor dem Einlesen des Datensatz wird aber empfohlen. Ebenfalls wird empfohlen das R-Package `lubridate` zu Laden um effizient mit der Variable `Datum` (Siehe Abschnitt 2.1.4) arbeiten zu können. Alle Code-Beispiele sind in R geschrieben und setzen ebenfalls das Package `tidyverse` bzw. dessen Bestandteile `dplyr`, `forcats` und `stringr` voraus.

1.3 Beteiligte

Die Qualitätskontrolle wurde unterstützt durch Johannes Schütt und Carla Mößner.

Die Erstellung des Codierschemas sowie die Umsetzung der Codierung wurden von Jakob Tures und den TeilnehmerInnen eines Lehrforschungsprojekts an der Universität Potsdam vom SoSe 2022 bis WiSe 2022/23 durchgeführt. In alphabetischer Reihenfolge: Sören Freisem, Maitilda Massa Gahein-Sama, Jakob Gustavs, Jacob Heise, Anton Gabor Gustav Leue, Andreas Niedlich und Josephine Penno.

Das vorliegende Datenhandbuch wurde von Nadja Dautel kontrolliert und editiert.

2 Variablen

Im Folgenden werden die im Datensatz enthaltenen Variablen beschrieben. Diese lassen sich dabei inhaltlich in mehrere Gruppen einteilen.

⁹Die Stammdaten der MdB wurden unter <https://www.bundestag.de/services/opendata> als XML heruntergeladen und ebenfalls in einen R-Datensatz überführt. Genutzt wurde die Datei mit dem Stand vom 19.09.2022.

¹⁰Neben dem Autor, waren dies die sieben TeilnehmerInnen des Lehrforschungsprojekts, siehe Abschnitt 1.3.

¹¹Zu den leichten Ungleichgewichten in der Umsetzung, siehe Abschnitt 2.3.1.

¹²In den Protokollen des Bundestages sind auch Zwischenrufe notiert, welche für die StenographInnen nicht verständlich waren (bspw. in der Form: „Zuruf der SPD.“) oder welche zwar verständlich waren, aber durch die StenographInnen nicht einer Fraktion und/oder MdB zugeordnet werden konnten.

¹³Während sich auf die Rate der False Negatives, also der Zwischenrufe die nicht erfolgreich aus den Protokollen extrahiert werden konnten, keine Angabe machen lässt, kann man die Rate der False Positives, also der fälschlicherweise als Zwischenruf extrahierten Protokollelemente, somit auf ca. 0,000056% schätzen.

2.1 Beobachtungsbezogen

Die Beobachtungen in diesem Datensatz sind alle 824,996 Zwischenreaktionen in den Protokollen des 17.–19. Bundestages. Die folgenden Variablen identifizieren die einzelnen Zwischenreaktionen und ermöglichen die zeitliche Zuordnung.

2.1.1 ID

Typ: Character

ID identifiziert jede Beobachtung des Datensatz durch einen einzigartigen Identifikationsstring. Die ersten zwei Stellen des Strings benennen die Legislaturperiode, gefolgt von drei Stellen welche die Sitzungsnummer innerhalb der Legislaturperiode angeben. Nach dem ersten Underscore wird die Redenposition innerhalb der Sitzung beziffert. Die Zahl nach dem zweiten Underscore beschreibt die Position innerhalb der Rede. So wurde jede Rede in Redebestandteil, Zwischenreaktion(en) auf diesen Redebestandteil, nächster Redebestandteil, Zwischenreaktion(en) auf diesen Redebestandteil, usw. aufgeteilt. Lange Reden mit vielen Zwischenreaktionen können so >100 Bestandteile haben.

So benennt bspw. die ID: 18042_17_32 den 32. Bestandteil der 17. Rede in der 42. Sitzung der 18. Legislaturperiode.

2.1.2 Periode

Typ: Integer

Periode beziffert die Legislaturperiode der Beobachtung.

Mögliche Werte sind: 17, 18, 19.

Periode	Anzahl	%
17	286,888	34.77
18	217,677	26.39
19	320,431	38.84

Tabelle 1: Zwischenreaktionen nach Legislaturperiode

2.1.3 Sitzung

Typ: Integer

Sitzung beziffert die Sitzungsnummer innerhalb einer Legislaturperiode.

Die 150. Sitzung der 17. Legislaturperiode fehlt im Datensatz. Diese ließ sich aufgrund von für diese PDF Datei spezifischen Problemen bei der oben beschriebenen Umwandlung in Text über OCR nicht in den Datensatz aufnehmen. Dass diese Sitzung ein Missing darstellt, sollte auf substantielle Analyseergebnisse keinen oder nur einen sehr kleinen Einfluss haben.

Periode	Anzahl Sitzungen	In Datensatz
17	253	252
18	245	245
19	239	239

Tabelle 2: Sitzungen nach Legislaturperiode

2.1.4 Datum

Typ: *Date*¹⁴

Datum gibt das Datum der Sitzung im Format YYYY-MM-DD an. Zur vollen Nutzung der Variable sollte das *lubridate* Package in R geladen werden. Mehr zu den Möglichkeiten von *lubridate* findet sich in *Wickham, Hadley & Grolemund, Garrett (2017) R for Data Science. Online-Ausgabe.* unter: <https://r4ds.had.co.nz/dates-and-times.html>.

2.2 Zwischenreaktion

Die folgenden Variablen unterscheiden zwischen Zwischenrufen und *weiteren Reaktionen* und enthalten den Text des Zwischenrufs, wenn vorhanden, bzw. die Art der Reaktion.

2.2.1 Typ

Typ: *Factor*

Typ unterscheidet nach Art der Reaktion. Das Level *Zwischenruf* deklariert eine Beobachtung als *Zwischenruf*, *Reaktion* als eine der *weiteren Reaktionen*: *Beifall*, *Heiterkeit*, *Lachen* und *Widerspruch* (siehe Abschnitt 2.2.3).

Typ	Anzahl	%
Reaktion	571,986	69.33
Zwischenruf	253,010	30.67

Tabelle 3: Zwischenreaktionen nach Typ

2.2.2 Zwischenruf

Typ: *Character*

Zwischenruf enthält den Textinhalt der Zwischenrufe. Bei Zwischenrufen mit Text, d.h. wenn die StenographInnen den Wortlaut notieren konnten, enthält die Variable den gesamten gesprochenen Text des Zwischenrufs als String. Bei Zwischenrufen ohne Text, d.h. wenn die StenographInnen den Wortlaut *nicht* notieren konnten, ist die Variable == *NA*.

Text	Anzahl	%
NA	13,488	5.33
Vorhanden	239,522	94.67

Tabelle 4: Zwischenrufe mit/ohne Text an allen *Zwischenrufen*

Zwischenruf ist ebenfalls == *NA*, wenn es sich nicht um einen Zwischenruf sondern eine der *weiteren Reaktionen* (siehe Abschnitt 2.2.3) handelt.

Text	Anzahl	%
NA	585,474	70.97
Vorhanden	239,522	29.03

Tabelle 5: Zwischenrufe mit/ohne Text an allen *Zwischenreaktionen*

¹⁴Klasse aus dem R-Package *lubridate*.

2.2.3 Reaktion

Typ: Factor

Reaktion unterscheidet die *weiteren Reaktionen* durch MdB oder Fraktionen nach ihrer Art. Zu den möglichen Levels: Beifall beschreibt Beifall und Applaus. Heiterkeit beschreibt wohlwollende Heiterkeit. Lachen beschreibt hämisches Lachen. Widerspruch umfasst nicht weiter beschriebene Widersprüche. Reaktion ist == NA wenn es sich nicht um eine der *weiteren Reaktionen* handelt, sondern ein Zwischenruf vorliegt.

Reaktion	Anzahl	%
Beifall	544,564	95.21
Heiterkeit	13,804	2.41
Lachen	8,839	1.55
Widerspruch	4,779	0.84

Tabelle 6: Weitere Reaktionen nach Typ

Je nach Ausrichtung einer Analyse, könnte entschieden werden, dass Widerspruch als Zwischenruf ohne Text interpretiert wird. Hierzu könnte der Datensatz wie folgt gefiltert werden:

```
bt_reactions %>%  
  filter(Typ == "Zwischenruf" | Reaktion == "Widerspruch")
```

2.3 Codierung

Die folgenden Variablen enthalten die manuelle Codierung des Inhalts von 17,999 Zwischenrufen in den Kategorien *Negativität*, *Inzivilität*, *Sexismus* und *Bezug* (inhaltlich oder persönlich) sowie eine Filtervariable zu deren Auswahl. Siehe dazu auch Abschnitte 1.2 und 2.3.1. Jede Kategorie wird im Folgenden kurz inhaltlich beschrieben. Für eine ausführlichere Darlegung ist das Codierschema¹⁵ zu beachten.

2.3.1 Sampled

Typ: Logical

Sampled kann genutzt werden um alle 17,999 Zwischenrufe auszuwählen für die eine Codierung des Inhalts (siehe Abschnitt 2.3) vorliegt. Das eigentlich angestrebte N von 18,000 Zwischenrufen konnte aufgrund eines False Positives, also einer als Zwischenruf identifizierten Beobachtung bei der es sich aber nicht um einen Zwischenruf handelte, nicht realisiert werden. Für alle Beobachtungen im Sample liegt der Wortlaut der Zwischenrufe in der Variable Zwischenruf als String vor. Außerdem ist für alle diese Beobachtungen Fraktion != NA und != "fraktionslos".

Sampled	Anzahl	%
FALSE	235,011	92.89
TRUE	17,999	7.11

Tabelle 7: Codierte Stichprobe an allen Zwischenrufen

Die Stichprobe wurde so geschichtet, dass pro Legislaturperiode 6.000 Zwischenrufe und innerhalb einer Legislaturperiode die gleiche Anzahl von Zwischenrufen pro Fraktion zufällig gezogen wurden. Die final realisierte Anzahl von Zwischenrufen pro Periode und Fraktion weichen leicht von diesen Zielgrößen ab (siehe Tabelle 8). Dies ist Erstens darauf zurückzuführen, dass ein Element der Stichprobe (aus der 17. Legislaturperiode) erst in der Codierung als False Positive erkennbar wurde, also kein Zwischenruf, sondern ein fehlerhaftes Artefakt des OCR-Prozess war. Zweitens hat sich aufgrund nachträglicher Fehlerkorrekturen die Fraktionszugehörigkeit für

¹⁵Codierschema_final.pdf

vier Beobachtungen der Stichprobe verändert.

Periode	Fraktion	Anzahl
17	CDU/CSU	1,200
17	FDP	1,201
17	BÜNDNIS 90/DIE GRÜNEN	1,200
17	SPD	1,199
17	DIE LINKE	1,199
18	CDU/CSU	1,500
18	BÜNDNIS 90/DIE GRÜNEN	1,500
18	SPD	1,500
18	DIE LINKE	1,500
19	CDU/CSU	998
19	FDP	1,000
19	BÜNDNIS 90/DIE GRÜNEN	1,000
19	SPD	1,000
19	DIE LINKE	999
19	AfD	1,003

Tabelle 8: Codierte Stichprobe nach Legislaturperiode und Fraktion

2.3.2 Negativ

Typ: Logical

Negativ ist für alle Zwischenrufe TRUE, für die vorliegende Negativität codiert wurde.

Negativität wurde dann codiert, wenn eine negative Bewertungsrichtung eines politischen Objekts (PolitikerIn, Partei, Institution) und/oder inhaltlicher Aspekte vorlag. Dies umfasst Sprache, welche sich kritisch, widersprechend oder abwertend auf ein politisches Objekt oder inhaltliche Aspekte bezieht. Negativität beginnt somit bereits bei sachlichem Widerspruch bzw. sachlicher Kritik, umfasst aber auch alle schwerwiegenden Formen von Negativität. Entsprechend wurden per Definition auch alle als inzivil sowie sexistisch codierten Zwischenrufe als negativ codiert.

Für die Codierung als negativ musste der Zwischenruf selbst Sprache enthalten die eine negative Bewertungsrichtung anzeigt. Ein Zwischenruf der einem negativen Punkt in einer Rede zustimmt, wurde demnach selbst nicht als negativ codiert.

Da die Codierung ausschließlich auf der geschriebenen Schriftsprache basierte, konnten Ironie und Sarkasmus nicht zuverlässig als solches erkannt werden. Auch hier war einzig die negative Bewertungsrichtung in der Sprache des Zwischenrufs entscheidend. Positive Sprache unter dem Verdacht der Ironie wurde somit nicht als Negativität codiert.

Negativ	Anzahl	%
FALSE	8,445	46.92
TRUE	9,554	53.08

Tabelle 9: Coding: *Negativ* an codierter Stichprobe

Fleiss Kappa¹⁶: 0,508

2.3.3 Sexism

Typ: Logical

¹⁶Im Rahmen der manuellen Codierung, wurden 100 Zwischenrufe durch alle 8 CodiererInnen bewertet. Zur Bewertung der Interdecoderreliabilität, wurde auf Basis dieser 100 Zwischenruf Fleiss Kappa berechnet. Die Berechnung erfolgte mit der R-Funktion `irr::kappam.fleiss(exact = FALSE)`.

Sexism ist für alle Zwischenrufe TRUE, für die vorliegender Sexismus codiert wurde.

Da subtile Formen von Alltagssexismus nicht objektiv codierbar waren, beschränkt sich die Kategorie in erster Linie auf offenen Sexismus in der Form von *Chauvinität*. Dieser war erkennbar an der Bezugnahme auf das Äußere (Aussehen, Kleidung, Stimme, ...) einer MdB, der Bezugnahme auf weibliche Rollenklischees sowie der Nutzung chauvinistischer Sprache. Darüber hinaus wurde auch bei einer expliziten Verknüpfung des Inhalts des Zwischenrufs mit dem weiblichen Geschlecht (bspw. „*Sie als Frau ...*“) Sexismus codiert.

Während der Codierung waren den CodiererInnen keine Informationen zu den ZwischenrufferInnen sowie den unterbrochenen RednerInnen bekannt, um eine Beeinflussung der Codierung zu vermeiden. Aus diesem Grund musste die Sexismus Kategorie nachträglich bereinigt werden. So behielt die Codierung als TRUE nur Bestand, wenn tatsächlich eine Frau durch einen Mann unterbrochen wurde. Da sich die zuverlässige Codierung dieser Kategorie als besonders schwierig herausstellte, wurden alle als sexistisch codierten Zwischenrufe nochmals in einer Gruppendiskussion mit allen CodiererInnen erörtert und auf Basis dessen teilweise korrigiert.

Sexism	Anzahl	%
FALSE	17,995	99.98
TRUE	4	0.02

Tabelle 10: Coding: *Sexism* an codierter Stichprobe

Aufgrund der extrem kleinen Anzahl so als sexistisch identifizierter Zwischenrufe, entzieht sich die Kategorie der statistischen Analyse. Die entsprechend codierten Zwischenrufe können jedoch als Beispiele für offene Formen von Sexismus im Deutschen Bundestag genutzt werden.

Fleiss Kappa: NA¹⁷

2.3.4 Incivility

Typ: Logical

Incivility ist für alle Zwischenrufe TRUE, für die vorliegende Inzivilität codiert wurde.

Unter Inzivilität im Sinne des Codierschemas wurde die Verletzung gesellschaftlich akzeptierter interpersonaler Kommunikations- und Umgangsnormen sowie Verstöße gegen das Ideal des fairen demokratischen Diskurs verstanden. Darunter fielen Respektlosigkeiten, „lustig machen“ über die RednerIn, Beschimpfungen, Beleidigungen, vulgäre Sprache, Bedrohungen, Bezichtigungen der Lüge, Rufmord, Unterstellungen, Attacken auf die Identität der RednerIn, Dämonisierungen und stereotype Feindbilder, Angriffe auf die Menschenwürde sowie starke Übertreibungen¹⁸. Ebenfalls als inzivil definiert war die Ausgrenzung aus dem demokratischen Diskurs bzw. der Parlamentsarbeit, unter anderem dem Absprechen der Fähigkeit die Rolle als MdB ausfüllen zu können oder zu dürfen sowie der Relevanz der Meinung einzelner oder ganzer Fraktionen/Parteien für den demokratischen Diskurs.

Da die Codierung ausschließlich auf der geschriebenen Schriftsprache basierte, konnten Ironie und Sarkasmus nicht zuverlässig als solches erkannt werden. Auch hier war einzig der inzivile Gehalt in der Sprache des Zwischenrufs entscheidend. Ein Zwischenruf der keine der oben beschriebenen Indikatoren vorwies wurde nicht als inzivil codiert, auch wenn möglicherweise Ironie oder Sarkasmus vermutbar waren.

Ein inziviler Zwischenruf war per Definition gleichzeitig auch als negativ zu codieren. Inzivilität ist stets auch negativ, Negativität wird aber erst zu Inzivilität, wenn allgemeine Kommunikationsnormen überschritten oder ignoriert werden. Zwischenrufe die als sexistisch codiert wurden mussten ebenfalls als inzivil codiert werden.

¹⁷Da in den 100 durch alle CodiererInnen codierten Zwischenrufe kein sexistischer Zwischenruf enthalten war, konnte Fleiss Kappa nicht berechnet werden, wäre aufgrund der extrem geringen Anzahl als TRUE codierter Zwischenrufe aber auch wenig aussagekräftig.

¹⁸Für eine ausführlichere Definition der gelisteten Unterpunkte, siehe das Codierschema *Codierschema_final.pdf*.

Incivility	Anzahl	%
FALSE	15,631	86.84
TRUE	2,368	13.16

Tabelle 11: Coding: *Incivility* an codierter Stichprobe

Fleiss Kappa: 0,635

2.3.5 Bezug

Typ: Factor

Bezug wurde in drei Leveln codiert. Als *Inhalt* codierte Zwischenrufe beziehen sich in ihrem Hauptinformationsgehalt auf den in der Rede behandelten Sachverhalt. *Person* zeigt an, dass der Hauptinformationsgehalt des Zwischenrufs sich auf die gerade sprechende RednerIn oder deren Fraktion bezieht. Bezüge auf andere Personen wurden als *Inhalt* codiert. *OB* (Ohne Bezug) diente als Auffangkategorie für unentscheidbare Bezüge. Diese durfte nur genutzt werden wenn Indikatoren für den inhaltlichen oder persönlichen Bezug gänzlich fehlten, was vor allem bei sehr kurzen Zwischenrufe mit ambivalenten/inhaltsleeren Begriffen wie „Oh!“ oder „Ah!“ der Fall sein konnte. Sobald Indikatoren für den inhaltlichen oder persönlichen Bezug vorlagen, musste entschieden werden, welcher Bezug überwog.

Bezug	Anzahl	%
Inhalt	13,710	76.17
Person	3,598	19.99
OB	691	3.84

Tabelle 12: Coding: *Bezug* an codierter Stichprobe

Fleiss Kappa: 0,416

Fleiss Kappa nach Kategorien:

Inhalt: 0,460

Person: 0,450

OB: 0,229

2.4 ZwischenruferIn

Die folgenden Variablen identifizieren das unterbrechende MdB, dessen Fraktion sowie Funktionen innerhalb der Regierung und der eigenen Fraktion sowie die aus den Stammdaten des Bundestages verfügbaren demographischen Informationen.

2.4.1 Name

Typ: Character

Name identifiziert das unterbrechende MdB namentlich.

In den Protokollen des Bundestages werden Zwischenreaktionen einzelner MdB namentlich notiert, wenn dies den StenographInnen möglich war. Kollektive Zwischenreaktionen einer ganzen oder großer Teile einer Fraktion (bspw. „Beifall bei der SPD“ oder „Zurufe vom BÜNDNIS 90/DIE GRÜNEN“) werden ohne Namen notiert. Für diese ist *Name* == NA.

Zwischenrufe werden in den Protokollen ebenfalls mit Namen notiert. Zwischenrufe die durch die StenographInnen nicht namentlich zugeordnet werden konnten, werden trotzdem (mit oder ohne Wortlaut) notiert. Für diese ist *Name* == NA.

Typ	Name	Anzahl	% in Typ
Reaktion	NA	541,318	94.64
Reaktion	Vorhanden	30,668	5.36
Zwischenruf	NA	22,772	9.00
Zwischenruf	Vorhanden	230,238	91.00

Tabelle 13: Zwischenreaktionen mit/ohne bekanntem Namen nach Typ der Zwischenreaktion

2.4.2 Fraktion

Typ: Factor

`Fraktion` identifiziert die Fraktion des unterbrechenden MdB.

In den Protokollen des Bundestages wird für Zwischenreaktionen die Fraktion des unterbrechenden MdB oder ganzer bzw. großer Teile einer Fraktion sowie mehrerer Fraktionen notiert, wenn dies den StenographInnen möglich war. Allgemeine Zwischenreaktionen des kompletten Plenums (bspw. „Heiterkeit im ganzen Hause“) werden ohne Fraktion notiert. Für diese ist `Fraktion == NA`.

Zwischenrufe werden in den Protokollen ebenfalls mit Fraktionszugehörigkeit notiert. Zwischenrufe die durch die StenographInnen nicht einer Fraktion zugeordnet werden konnten, werden trotzdem (mit oder ohne Wortlaut) notiert. Für diese ist `Fraktion == NA`.

Typ	Fraktion	Anzahl	% in Typ
Reaktion	NA	5,661	0.99
Reaktion	Vorhanden	566,325	99.01
Zwischenruf	NA	1,220	0.48
Zwischenruf	Vorhanden	251,790	99.52

Tabelle 14: Zwischenreaktionen mit/ohne bekannter Fraktion nach Typ der Zwischenreaktion

`Fraktion` kann als Level alle Fraktionen des 17.-19. Deutschen Bundestag, sowie `fraktionslos` für fraktionslose Abgeordnete, annehmen

Periode	Fraktion	Anzahl	% in Periode
17	CDU/CSU	68,162	23.76
17	FDP	59,848	20.86
17	BÜNDNIS 90/DIE GRÜNEN	47,281	16.48
17	SPD	70,969	24.74
17	DIE LINKE	38,258	13.34
17	fraktionslos	3	0.00
17	NA	2,367	0.83
18	CDU/CSU	63,999	29.40
18	BÜNDNIS 90/DIE GRÜNEN	52,234	24.00
18	SPD	55,135	25.33
18	DIE LINKE	43,411	19.94
18	fraktionslos	5	0.00
18	NA	2,893	1.33
19	CDU/CSU	64,955	20.27
19	FDP	44,817	13.99
19	BÜNDNIS 90/DIE GRÜNEN	53,336	16.65
19	SPD	65,046	20.30
19	DIE LINKE	47,406	14.79
19	fraktionslos	138	0.04
19	AfD	43,112	13.45
19	NA	1,621	0.51

Tabelle 15: Zwischenreaktionen nach Fraktion und Legislaturperiode

2.4.3 Regierungsfunktion

Typ: Character

`Regierungsfunktion` beschreibt welche Regierungsrolle im weiteren Sinne das unterbrechende MdB ausübt. Der Name der Rolle liegt als String vor. Neben der KanzlerIn und den MinisterInnen sind auch Parlamentarische StaatssekretärInnen, BundespräsidentInnen sowie MinisterInnen und SenatorInnen auf Landesebene als Regierungsfunktionen gelistet. Diese sprechen als Mitglieder der Exekutive, des Bundesrats oder als Gäste im Bundestag. Entsprechend, können sie auch Reden unterbrechen, was allerdings nur sehr selten vorkommt (siehe Tabelle 16).

Die Regierungsfunktionen wurden aus den Protokollen so extrahiert wie sie dort notiert waren und in Einzelfällen manuell korrigiert.

Für ein unterbrechendes MdB ohne Regierungsfunktion, wie sie in den Protokollen notiert war, ist `Regierungsfunktion == NA`. Bei fehlenden Informationen zu dem Namen des MdB ist die Regierungsfunktion notwendigerweise ebenfalls NA.

Typ	Regierungsfunktion	Anzahl	% in Typ
Reaktion	NA	571,986	100.00
Zwischenruf	NA	252,509	99.80
Zwischenruf	Vorhanden	501	0.20

Tabelle 16: Zwischenreaktionen mit/ohne bekannter Regierungsfunktion nach Typ der Zwischenreaktion

Analysen mit dieser extrem kleinen Gruppe sind für die meisten Zwecke vermutlich nicht sinnvoll. Soll die Variable trotzdem genutzt werden, bietet sich ein Preprocessing wie unter Abschnitt 2.6.4 beschrieben an.

2.4.4 Fraktionsfunktion

Typ: Character

`Fraktionsfunktion` beschreibt welche Funktion das unterbrechende MdB in dessen Fraktion ausübt. Der Name der Funktion(en) liegt als String vor. Dies sind unter anderem (stellvertretende) Vorsitzende, Parlamentarische GeschäftsführerInnen, JustiziarInnen aber auch speziellere Rollen.

Die Funktionen innerhalb der Fraktionen wurden aus den Stammdaten des MdB extrahiert und den Personen anhand der Namen und der Legislaturperiode zugespielt. Die Stammdaten enthalten detaillierte zeitliche Informationen dazu, wann genau ein MdB eine Funktion ausgefüllt hat. Dies wurde in so fern vereinfacht, als dass die Variable für jede Legislaturperiode anzeigt, welche Funktion(en) das MdB in dieser Periode insgesamt ausgefüllt hat, unabhängig davon, ob dies auch am Tage der Sitzung tatsächlich der Fall war.

Für ein unterbrechendes MdB ohne `Fraktionsfunktion`, wie sie in den Stammdaten aufgeführt waren, ist `Fraktionsfunktion == NA`. Bei fehlenden Informationen zu dem Namen des MdB ist die `Fraktionsfunktion` notwendigerweise ebenfalls NA.

Typ	Fraktionsfunktion	Anzahl	% in Typ
Reaktion	NA	564,730	98.73
Reaktion	Vorhanden	7,256	1.27
Zwischenruf	NA	160,369	63.38
Zwischenruf	Vorhanden	92,641	36.62

Tabelle 17: Zwischenreaktionen mit/ohne bekannter `Fraktionsfunktion` nach Typ der Zwischenreaktion

Für Analysen, sollte die Variable in den meisten Fällen noch weiter preprocessed werden. So kann beispielsweise eine Dummy-Variable gebildet werden, welche das Vorliegen/Fehlen einer `Fraktionsfunktion` anzeigt:

```
bt_reactions %>%
  mutate(Funktion = if_else(
    !is.na(Fraktionsfunktion),
    TRUE, FALSE
  ))
```

Je nach Fragestellung ist auch ein differenzierterer Umgang mit den unterschiedlichen Funktionen denkbar. Beispielsweise unter Nutzung von *Regular Expressions* (siehe Abschnitt 2.6.4) oder mit einer manuellen Recodierung der Funktionen.

2.4.5 Gender

Typ: Factor

`Gender` zeigt das binäre Gender des unterbrechenden MdB an. Die möglichen Levels sind weiblich und männlich.

Das Gender wurde aus den Stammdaten des MdB extrahiert und den Personen anhand der Namen zugespielt. Für alle MdB die in den Stammdaten geführt werden ist das Gender bekannt. `Gender` ist somit nur für Beobachtungen NA für welche die Zuordnung der Stammdaten nicht möglich war. Dies ist vor allem der Fall, wenn `Name == NA` ist¹⁹, aber auch für unterbrechende Personen die kein MdB sind und somit nicht in den Stammdaten geführt werden.

¹⁹Weil die StenographInnen den Namen der unterbrechenden MdB nicht notieren konnten oder weil eine Zwischenreaktion/Zwischenruf einer ganzen oder großen Teilen einer Fraktion zugeschrieben wurde.

Typ	Gender	Anzahl	% in Typ
Reaktion	weiblich	12,122	2.12
Reaktion	männlich	18,532	3.24
Reaktion	NA	541,332	94.64
Zwischenruf	weiblich	72,599	28.69
Zwischenruf	männlich	157,551	62.27
Zwischenruf	NA	22,860	9.04

Tabelle 18: Zwischenreaktionen nach Gender und Typ der Zwischenreaktion

2.4.6 Alter

Typ: Integer

`Alter` zeigt das Alter des unterbrechenden MdB zu Beginn der jeweiligen Legislaturperiode an.

Das Geburtsdatum des MdB wurde aus den Stammdaten des MdB extrahiert und auf Basis dieser das Alter zu Beginn der jeweiligen Legislaturperioden berechnet. Für die 17. Legislaturperiode war dies der 27. Oktober 2009, für die 18. der 22. Oktober 2013 und für die 19. der 24. Oktober 2017. Das so berechnete Alter wurde den Personen anhand der Namen zugespielt.

Für alle MdB die in den Stammdaten geführt werden, ist das Alter bzw. Geburtsdatum bekannt. `Alter` ist somit nur für Beobachtungen == NA für welche die Zuordnung der Stammdaten nicht möglich war. Dies ist vor allem der Fall, wenn `Name == NA` ist²⁰, aber auch für unterbrechende Personen die keine MdB sind und somit nicht in den Stammdaten geführt werden.

Typ	Alter	Anzahl	% in Typ
Reaktion	NA	541,332	94.64
Reaktion	Vorhanden	30,654	5.36
Zwischenruf	NA	22,860	9.04
Zwischenruf	Vorhanden	230,150	90.96

Tabelle 19: Zwischenreaktionen mit/ohne bekanntem Alter nach Typ der Zwischenreaktion

Typ	P0	P25	P50	P75	p100	Mittelwert
Reaktion	22	42.00	51.00	57.00	77	50.02
Zwischenruf	22	42.00	50.00	56.00	77	48.87

Tabelle 20: Altersverteilung nach Typ der Zwischenreaktion

2.4.7 Akad_Titel

Typ: Character

`Akad_Titel` beschreibt den akademischen Titel des unterbrechenden MdB als String.

Die akademischen Titel wurden zunächst aus den Protokollen ausgelesen, in welchen sie im Normalfall mit dem Namen des unterbrechenden MdB notiert sind. Anhand der Informationen zu akademischen Titeln aus den Stammdaten des MdB konnten in einem zweiten Schritt Fehler korrigiert und Missings gefüllt werden.

Für unterbrechende MdB ohne akademischen Titel, wie sie in den Protokollen und Stammdaten aufgeführt waren, ist `Akad_Titel == NA`. Bei fehlenden Informationen zu dem Namen der MdB ist `Akad_Titel` notwendigerweise ebenfalls == NA.

²⁰Siehe Fußnote 19

Typ	Akad_Titel	Anzahl	% in Typ
Reaktion	NA	565,267	98.83
Reaktion	Vorhanden	6,719	1.17
Zwischenruf	NA	209,679	82.87
Zwischenruf	Vorhanden	43,331	17.13

Tabelle 21: Zwischenreaktionen mit/ohne bekanntem akademischen Titel nach Typ der Zwischenreaktion

Für Analysen, sollte die Variable in den meisten Fällen noch weiter preprocessed werden. So kann beispielsweise eine Dummy-Variable gebildet werden, welche das Vorliegen/Fehlen eines akademischen Titels anzeigt:

```
bt_reactions %>%
  mutate(Titel = if_else(
    !is.na(Akad_Titel),
    TRUE, FALSE
  ))
```

Je nach Fragestellung ist auch ein differenzierterer Umgang mit den unterschiedlichen Titeln denkbar, bspw. die Unterscheidung zwischen DoktorInnen und ProfessorInnen oder Ehrentiteln (h. c.) und regulären akademischen Titeln. Dies lässt sich bspw. mit *Regular Expressions* (siehe Abschnitt 2.6.4) oder der manuellen Recodierung umsetzen.

2.5 Fraktionsbezogen

Die folgenden Variablen beziehen sich auf die Fraktion des unterbrechenden MdB. Die Werte sind somit für alle MdB einer Fraktion innerhalb einer Legislaturperiode konstant.

2.5.1 Regierungsfraktion

Typ: Logical

Regierungsfraktion identifiziert ob eine Fraktion in einer Legislaturperiode an der Regierung beteiligt war. Dies zeigt damit nicht an, dass das MdB eine Rolle in der Regierung inne hatte, sondern ob ihre Fraktion Teil der jeweiligen Regierungskoalition war. MdB welche in ihrer Regierungsfunktion sprechen oder unterbrechen, haben stets `Fraktion == NA`. Dies wird in den Protokollen dadurch reflektiert, dass für diese Personen keine Fraktionszugehörigkeit, aber der Name der Regierungsfunktion notiert wird, und wurde entsprechend in den Datensatz übernommen. Für Informationen zu der Regierungsrolle des MdB siehe Abschnitt 2.4.3.

Die Variable wurde manuell erstellt und dem Datensatz anhand Fraktion und Legislaturperiode zugespielt.

Für Beobachtungen mit `Fraktion == NA` ist Regierungsfraktion notwendigerweise NA.

Typ	Regierungsfraktion	Anzahl	%
Reaktion	FALSE	278,010	48.60
Reaktion	TRUE	288,315	50.41
Reaktion	NA	5,661	0.99
Zwischenruf	FALSE	162,960	64.41
Zwischenruf	TRUE	88,830	35.11
Zwischenruf	NA	1,220	0.48

Tabelle 22: Zwischenreaktionen nach Regierungsfraktion und Typ der Zwischenreaktion

2.5.2 Fraktionsgroesse

Typ: Integer

Fraktionsgroesse identifiziert die Anzahl der MdB aus der eine Fraktion in einer Legislaturperiode bestand.

Die Anzahl der MdB pro Fraktion und Legislaturperiode wurde auf Basis der Stammdaten des Bundestages berechnet und dem Datensatz anhand Fraktion und Legislaturperiode zugespielt.

Für Beobachtungen mit `Fraktion == NA` ist Fraktionsgroesse notwendigerweise NA.

Typ	Fraktionsgroesse	Anzahl	% in Typ
Reaktion	NA	5,661	0.99
Reaktion	Vorhanden	566,325	99.01
Zwischenruf	NA	1,220	0.48
Zwischenruf	Vorhanden	251,790	99.52

Tabelle 23: Zwischenreaktionen mit/ohne bekannter Fraktionsgroesse nach Typ der Zwischenreaktion

Periode	Fraktion	Fraktionsgroesse
17	CDU/CSU	243
17	FDP	100
17	BÜNDNIS 90/DIE GRÜNEN	75
17	SPD	153
17	DIE LINKE	76
17	fraktionslos	5
18	CDU/CSU	322
18	BÜNDNIS 90/DIE GRÜNEN	64
18	SPD	203
18	DIE LINKE	66
18	fraktionslos	3
19	CDU/CSU	259
19	FDP	85
19	BÜNDNIS 90/DIE GRÜNEN	72
19	SPD	164
19	DIE LINKE	71
19	fraktionslos	13
19	AfD	86

Tabelle 24: Fraktionsgroesse nach Fraktion und Legislaturperiode

2.6 Reaktion auf

Die im Vorangegangenen beschriebenen personen- und fraktionsbezogenen Variablen liegen ebenfalls für die unterbrochenen RednerInnen vor. Die folgenden Variablen identifizieren und beschreiben die unterbrochene RednerIn sowie deren Fraktion.

2.6.1 Reacted_Name

Typ: Character

`Reacted_Name` identifiziert die unterbrochene RednerIn namentlich.

In den Protokollen des Bundestages werden die Namen aller RednerInnen notiert. Der geringe Anteil von fehlenden Namen der RednerInnen in Tabelle 25 ist auf Fehler im OCR Prozess

zurückzuführen²¹.

Typ	Reacted_Name	Anzahl	% in Typ
Reaktion	NA	6,522	1.14
Reaktion	Vorhanden	565,464	98.86
Zwischenruf	NA	2,879	1.14
Zwischenruf	Vorhanden	250,131	98.86

Tabelle 25: Zwischenreaktionen mit/ohne bekanntem Namen *der RednerIn* nach Typ der Zwischenreaktion

2.6.2 Reacted_Fraktion

Typ: Factor

`Reacted_Fraktion` identifiziert die unterbrochene Fraktion als Faktorvariable.

In den Protokollen des Bundestages werden die Fraktionen aller RednerInnen notiert. Der Anteil von fehlenden Fraktionen der RednerInnen in Tabelle 26 ist auf Fehler im OCR Prozess sowie vor allem Reaktionen auf Reden von Regierungsmitgliedern (siehe Tabelle 29) und GastrednerInnen zurückzuführen.

Typ	Reacted_Fraktion	Anzahl	% in Typ
Reaktion	NA	109,063	19.07
Reaktion	Vorhanden	462,923	80.93
Zwischenruf	NA	30,182	11.93
Zwischenruf	Vorhanden	222,828	88.07

Tabelle 26: Zwischenreaktionen mit/ohne bekannter Fraktion *der RednerIn* nach Typ der Zwischenreaktion

²¹Da in den Protokollen jede Rede mit einem Namen versehen sein sollte, deutet dies auf eine Fehlerquote von ca. 1,14% für die ausgelesenen Reden hin. Neben Problemen die im OCR-Prozess oder in der Extraktion der Daten aus dem Fließtext mit Regular Expressions entstanden sein könnten, ist aus den Erfahrungen der manuellen Fehlerkorrekturen auch bekannt, dass die Plenarprotokolle sich nicht immer genau an die generell erkennbare Standardformatierung hielten. Dies führte regelmäßig zu weiteren Problemen. Für die ausgelesenen Zwischenreaktionen, lässt sich die Fehlerquote nicht schätzen. Da aber der Fokus der extensiven manuellen Fehlerkorrekturen auf der Sicherung der Qualität der Daten zu den Zwischenreaktionen lag, ist davon auszugehen, dass die Fehlerquote hier deutlich geringer ist. Ein Abgleich von 100 zufällig ausgewählten Zwischenrufen mit vorhandenem Text aus Version 1.0 des Datensatzes mit den PDF-Protokollen ließ für diese Subgruppe keine Fehler mehr erkennen. Weitere Qualitätskontrollen dieser Art und manuelle Fehlerkorrekturen, inklusive der Informationen zu den RednerInnen, sind für zukünftige Versionen des Datensatzes vorgesehen.

Periode	Reacted_Fraktion	Anzahl	% in Periode
17	CDU/CSU	76,044	26.51
17	SPD	56,227	19.60
17	FDP	42,807	14.92
17	DIE LINKE	28,093	9.79
17	BÜNDNIS 90/DIE GRÜNEN	37,451	13.05
17	NA	46,266	16.13
18	CDU/CSU	62,089	28.52
18	SPD	48,672	22.36
18	DIE LINKE	30,491	14.01
18	BÜNDNIS 90/DIE GRÜNEN	36,624	16.82
18	fraktionslos	15	0.01
18	NA	39,786	18.28
19	CDU/CSU	70,023	21.85
19	SPD	54,944	17.15
19	FDP	34,145	10.66
19	DIE LINKE	32,097	10.02
19	BÜNDNIS 90/DIE GRÜNEN	37,660	11.75
19	fraktionslos	615	0.19
19	AfD	37,754	11.78
19	NA	53,193	16.60

Tabelle 27: Zwischenreaktionen nach Fraktion *der RednerIn* und Legislaturperiode

2.6.3 Reacted_Regierungsfraktion

Typ: Logical

`Reacted_Regierungsfraktion` identifiziert, ob die unterbrochene Fraktion in einer Legislaturperiode an der Regierung beteiligt war. Dies zeigt damit nicht an, dass das unterbrochene MdB eine Rolle in der Regierung inne hatte sondern ob dessen Fraktion Teil der jeweiligen Regierungskoalition war. MdB mit Regierungsfunktion haben stets `Fraktion == NA`. Für Informationen zu der Regierungsrolle des unterbrochenen MdB siehe Abschnitt 2.6.4.

Die Variable wurde manuell erstellt und dem Datensatz anhand `Fraktion` und `Legislaturperiode` zugespielt.

Typ	Reacted_Regierungsfraktion	Anzahl	%
Reaktion	FALSE	231,360	40.45
Reaktion	TRUE	231,563	40.48
Reaktion	NA	109,063	19.07
Zwischenruf	FALSE	99,812	39.45
Zwischenruf	TRUE	123,016	48.62
Zwischenruf	NA	30,182	11.93

Tabelle 28: Zwischenreaktionen nach Regierungsfraktion *der RednerIn* und Typ der Zwischenreaktion

2.6.4 Reacted_Regierungsfunktion

Typ: Character

`Reacted_Regierungsfunktion` beschreibt welche Regierungsrolle im weiteren Sinne das unterbrochene MdB ausübt. Der Name der Rolle liegt als String vor. Neben der KanzlerIn und den MinisterInnen sind auch Parlamentarische StaatssekretärInnen, BundespräsidentInnen sowie MinisterInnen und SenatorInnen auf Landesebene als Regierungsfunktionen gelistet. Diese sprechen als Mitglieder der Exekutive, des Bundesrates oder als Gäste im Bundestag.

Die Regierungsfunktionen wurden aus den Protokollen extrahiert wie sie dort notiert waren und in Einzelfällen manuell korrigiert.

Für unterbrochene MdB ohne Regierungsfunktion, wie sie in den Protokollen notiert war, ist Regierungsfunktion == NA.

Typ	Reacted_Regierungsfunktion	Anzahl	% in Typ
Reaktion	NA	469,042	82.00
Reaktion	Vorhanden	102,944	18.00
Zwischenruf	NA	225,526	89.14
Zwischenruf	Vorhanden	27,484	10.86

Tabelle 29: Zwischenreaktionen mit/ohne bekannter Regierungsfunktion *der RednerIn* nach Typ der Zwischenreaktion

Je nach Fragestellung ist auch ein differenzierterer Umgang mit den unterschiedlichen Regierungsfunktionen denkbar. Dies kann mit einer manuellen Recodierung der Funktionen oder unter Einsatz von *Regular Expressions* realisiert werden. Um bspw. eine Dummy-Variable für alle Mitglieder der Exekutive im engeren Sinne, also Bundeskanzlerin und MinisterInnen zu erstellen:

```
bt_reactions %>%
  mutate(
    Exekutive = if_else(
      !is.na(Reacted_Regierungsfunktion) &
      str_detect(Reacted_Regierungsfunktion,
        pattern = "Bundeskanzlerin|^Bundesminister"),
      TRUE, FALSE
    )
  )
```

Vergleichbare Techniken könnten auch für die weiteren Character-Variablen mit vielen Ausprägungen im Datensatz angewandt werden (siehe insbesondere Abschnitte 2.4.3, 2.4.4 und 2.4.7).

2.6.5 Reacted_Gender

Typ: Factor

Reacted_Gender identifiziert das binäre Gender des unterbrochenen MdB. Die möglichen Levels sind weiblich und männlich.

Das Gender wurde aus den Stammdaten der MdB extrahiert und den Personen anhand der Namen zugespielt. Für alle MdB die in den Stammdaten geführt werden, ist das Gender bekannt. Gender ist somit nur für Beobachtungen NA für welche die Zuordnung der Stammdaten nicht möglich war, also in erster Linie für GastrednerInnen, welche keine MdB sind und somit nicht in den Stammdaten geführt werden, sowie für die wenigen RednerInnen mit fehlendem Namen (siehe Tabelle 25).

Typ	Reacted_Gender	Anzahl	% in Typ
Reaktion	männlich	361,761	63.25
Reaktion	weiblich	195,486	34.18
Reaktion	NA	14,739	2.58
Zwischenruf	männlich	176,653	69.82
Zwischenruf	weiblich	70,256	27.77
Zwischenruf	NA	6,101	2.41

Tabelle 30: Zwischenreaktionen nach Gender *der RednerIn* und Typ der Zwischenreaktion

2.7 Gewichte

Die Stichprobe der manuell codierten Zwischenrufe wurde als geschichtete Zufallsstichprobe gezogen. Pro Legislaturperiode wurden 6.000 Zwischenrufe und innerhalb einer Legislaturperiode die gleiche Anzahl von Zwischenrufen pro Fraktion zufällig gezogen. Dabei wurden nur Zwischenrufe gezogen, für welche sowohl der Text als auch die Fraktion (ohne *fraktionslose*) bekannt war. Um Analysen auf Basis der geschichteten Zufallsstichprobe verzerrungsfrei zu ermöglichen, wurden Gewichte berechnet und dem Datensatz hinzugefügt.

PW und IPW sind für alle Beobachtungen die nicht Teil der Stichprobe waren²² == NA.

2.7.1 PW

Typ: Double

PW enthält die Probability Weights der Auswahlwahrscheinlichkeit. Diese wurden pro Legislaturperiode und Fraktion berechnet als n/N , wobei n die Anzahl gezogener Zwischenrufe mit Text einer Fraktion und N alle Zwischenrufe mit Text einer Fraktion sind.

2.7.2 IPW

Typ: Double

IPW enthält die Inverse Probability Weights der Auswahlwahrscheinlichkeit. Diese wurden pro Legislaturperiode und Fraktion berechnet als $1/(n/N)$, wobei n die Anzahl gezogener Zwischenrufe mit Text einer Fraktion und N alle Zwischenrufe mit Text einer Fraktion sind.

Periode	Fraktion	PW	IPW
17	CDU/CSU	0.06	16.96
17	FDP	0.08	12.42
17	BÜNDNIS 90/DIE GRÜNEN	0.06	15.40
17	SPD	0.04	26.55
17	DIE LINKE	0.11	9.26
18	CDU/CSU	0.10	9.86
18	BÜNDNIS 90/DIE GRÜNEN	0.07	15.19
18	SPD	0.17	5.75
18	DIE LINKE	0.12	8.24
19	CDU/CSU	0.07	13.52
19	FDP	0.08	12.29
19	BÜNDNIS 90/DIE GRÜNEN	0.07	15.17
19	SPD	0.08	11.92
19	DIE LINKE	0.08	12.12
19	AfD	0.06	16.70

Tabelle 31: Gewichte nach Periode und Fraktion

3 Anwendung der Gewichte mit `survey`

Im folgenden wird beispielhaft auf die Anwendung der Gewichte zur Berechnung inferentieller Statistiken mit dem R-Package `survey` eingegangen.

Dieses Package kann zur Deklaration einer komplexen Stichprobe und zur nachfolgenden Berechnung unverzerrter Schätzer sowie Standardfehler genutzt werden. Als exemplarisches Analyseziel sollen zunächst die absoluten und relativen Anteile an negativen Zwischenrufen in der Grundgesamtheit aller Zwischenrufe des 19. Bundestages insgesamt und nach Fraktion geschätzt werden. Im Anschluss soll der Effekt der Fraktionszugehörigkeit auf die Wahrscheinlichkeit einen negativen Zwischenruf zu tätigen in einer logistischen Regression geschätzt werden.

²²Ermittelbar über `filter(Sampled == TRUE)`, siehe auch Abschnitt 2.3.1

Zunächst muss das Package geladen und der Datensatz vorbereitet werden. Dabei wird der Datensatz auf die Elemente der Stichprobe des 19. Bundestages reduziert und der leere Faktor `fraktionslos` entfernt.

```
library(survey)

bt_reactions_19 <- bt_reactions %>%
  filter(Sampled == TRUE & Periode == 19) %>%
  mutate(Fraktion = fct_drop(Fraktion, "fraktionslos"))
```

Um die geschichtete Stichprobe korrekt zu deklarieren, muss zunächst noch ein Identifier für die einzelnen Strata in denen gezogen wurde konstruiert werden. Dieser Identifier, die Probability Weights sowie der eigentliche Datensatz, werden in `svydesign` zur Deklaration der komplexen Stichprobe verwendet.

```
bt_reactions_19 <- bt_reactions_19 %>%
  mutate(Strats = str_c(Periode, Fraktion))

dsg <- svydesign(ids = ~ 0,
  strata = ~ Strats,
  fpc = ~ PW,
  data = bt_reactions_19)
```

Das so erstellte Objekt kann nun für gewichtete Analysen genutzt werden.

Die absolute Anzahl negativer und nicht-negativer Zwischenrufe in der Grundgesamtheit sowie der Standardfehler lassen sich mit `svytotal()` schätzen:

```
svytotal(~ Negativ, design = dsg)
```

	total	SE
NegativFALSE	35813.66	505.40
NegativTRUE	45909.26	505.40

Tabelle 32: Gewichtete Schätzung der absoluten (nicht-)negativen Zwischenrufe im 19. Bundestag

Analog für die relative Häufigkeit, mit `svymean()`:

```
svymean(~ Negativ, design = dsg)
```

	mean	SE
NegativFALSE	0.44	0.01
NegativTRUE	0.56	0.01

Tabelle 33: Gewichtete Schätzung des relativen Anteils (nicht-)negativer Zwischenrufe im 19. Bundestag

Eine Möglichkeit zur Schätzung der Werte in Abhängigkeit einer zweiten Variable ist `svyby`. `FUN = svytotal` fordert die absoluten Werte an, `FUN = svymean` die relativen:

```
svyby(~ Negativ, ~ Fraktion,
  design = dsg, FUN = svytotal)
```

Fraktion	NegativFALSE	NegativTRUE	se.NegativFALSE	se.NegativTRUE
CDU/CSU	6464.95	7033.00	205.51	205.51
FDP	6376.95	5910.05	186.16	186.16
BÜNDNIS 90/DIE GRÜNEN	6355.39	8812.61	228.84	228.84
SPD	5362.65	6554.35	179.53	179.53
DIE LINKE	5125.91	6979.97	181.36	181.36
AfD	6127.80	10619.29	247.09	247.09

Tabelle 34: Gewichtete Schätzung der absoluten (nicht-)negativen Zwischenrufe im 19. Bundestag nach Fraktion

```
svyby(~ Negativ, ~ Fraktion,
      design = dsg, FUN = svymean)
```

Fraktion	NegativFALSE	NegativTRUE	se.NegativFALSE	se.NegativTRUE
CDU/CSU	0.48	0.52	0.02	0.02
FDP	0.52	0.48	0.02	0.02
BÜNDNIS 90/DIE GRÜNEN	0.42	0.58	0.02	0.02
SPD	0.45	0.55	0.02	0.02
DIE LINKE	0.42	0.58	0.01	0.01
AfD	0.37	0.63	0.01	0.01

Tabelle 35: Gewichtete Schätzung des relativen Anteile (nicht-)negativer Zwischenrufe im 19. Bundestag nach Fraktion

Eine gewichtete logistische Regression der Wahrscheinlichkeit eines negativen Zwischenruf in Abhängigkeit der Fraktionszugehörigkeit lässt sich mit `svyglm()` berechnen:

```
svyglm(Negativ ~ Fraktion,
      design = dsg, family = binomial)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0842	0.0610	1.38	0.1675
FraktionFDP	-0.1603	0.0861	-1.86	0.0626
FraktionBÜNDNIS 90/DIE GRÜNEN	0.2427	0.0870	2.79	0.0053
FraktionSPD	0.1165	0.0862	1.35	0.1767
FraktionDIE LINKE	0.2245	0.0865	2.59	0.0095
FraktionAfD	0.4656	0.0881	5.28	0.0000

Tabelle 36: Logistische Regression der Negativität auf die Fraktionszugehörigkeit