Online Appendices:

Legislators' Emotional Engagement with Women's Issues: Gendered Patterns of Vocal Pitch in the German Bundestag

Oliver Rittmann[*]

[*] School of Social Sciences, University of Mannheim. E-mail: orittman@mail.uni-mannheim.de.

A The Representation of Women in the US House of Representatives and the German Bundestag over Time

To substantiate the discussion of the history of women's political representation in the US House of Representatives and the German Bundestag, figure A1 gives an overview of women's descriptive representation in both parliaments over time. Women remain underrepresented in both parliaments. Still, the figure shows that women have been represented at constantly higher rates in the German Bundestag than in the US House of Representatives. At the end of the study period of the US House-study in August 2014, the share of women in the US House of Representatives was 18.0%. During the same time, the share of women in Germany was more than 18 percentage points higher, at 36.3%.

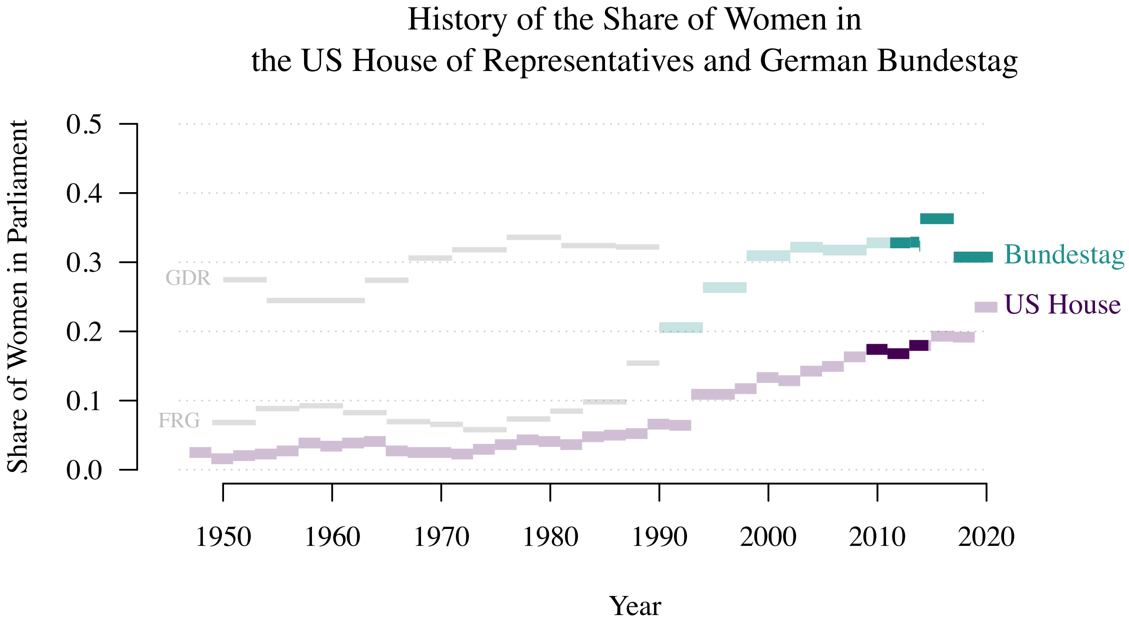

Figure A1: The figure shows how the share of women legislators in the US House of Representatives (1946–2020) and the German Bundestag (1990–2020) evolved over time. Prior to 1990, the figure also shows the share of women in the *Bundesrat* of the former German Federal Republic (FRG), and in the *Volkskammer* of the former German Democratic Republic (GDR). Dietrich, Hayes and O'Brien (2019) study speeches in the US House of Representatives between August 2009 and January 2014. The replication analysis in Germany considers speeches in the Bundestag between February 2011 and July 2020. Both periods are highlighted by solid colors in the graphic. Data source: Inter-Parliamentary Union (2022).

B Download and Pre-processing of Audio Data

All audio data used in the analysis was downloaded from the video archive of the German Bundestag, `www.bundestag.de/mediathek`. The archive is hierarchically organized on three levels: Plenary sessions, agenda items, and individual speeches. Figure A2 outlines the nested structure of the archive. Agenda items are nested within the plenary session of the day, and individual speeches are nested within the agenda item they belong to. The interface of the website allows users to access and download videos on either of the three levels. This means that users can either download the recording of an entire plenary session as one video, they can download all proceedings during one specific agenda item as one video, or they can download videos of individual speeches. Because the unit of analysis of the present study are individual speeches, I downloaded videos on the speech level along with relevant meta-data. This meta-data includes the name of the speaker, the date of the speech, and the agenda item the speech belongs to.

In total, I downloaded 50,198 videos of individual speeches from between February 23, 2011 and July 3, 2020. Because the analysis only makes use of the audio modality, I extract the audio track (`.wav`-format) from the video files (`mp4`-format) using the open-source software `FFmpeg`. The average audio file is 5.43 minutes long and has a file size of 62.24 MB. This yields a total of 4,542 hours of audio recordings, and a total file size of 3.12 TB.

All `.wav`-files were analyzed using the open-source software `Praat` (Boersma and Weenik, 2021). As explained in the manuscript, this software was used to extract the mean fundamental frequency F_0 as a measure of vocal pitch from the audio files. `Praat` requires users to set floor and ceiling parameters. Following Dietrich, Hayes and O'Brien (2019), I set those parameters to 100 Hz and 500 Hz for women and to 75 Hz and 300 Hz for men. Next, I center the F_0 scores around legislator-specific means, and scale it to reflect standard deviation from legislator-specific means. This strategy follows Dietrich, Hayes and O'Brien (2019) and adjusts for the fact that different legislators speak at a generally higher or lower pitch. Figure A3 visualizes this procedure. The left panel shows

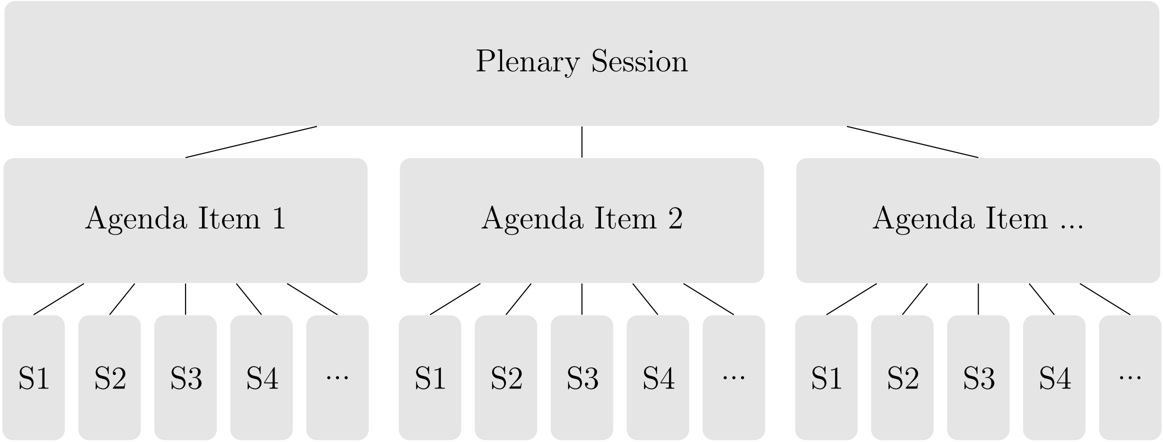

Figure A2: Structure of `https://www.bundestag.de/mediathek` for one session day.

the distribution of average fundamental frequencies of all parliamentary speeches in the data set. It stands out that women's and men's average fundamental frequencies follow distinct distributions with women speaking at a higher vocal pitch than men. This difference disappears after centering the distribution around legislator-specific averages (center panel). After dividing the distribution by legislator-specific standard deviation, the distribution of (now z-standardized) fundamental frequency reflects standard deviations from the speaker's average level of vocal pitch.

It is important to note that audio files were not cleaned with respect to disturbing noises (e.g. applause), or infrequent occurrences of other speakers. For example, videos usually start with a few words spoken by the chair to introduce the speaker. Moreover, the audio files sometimes include utterances by other speakers, e.g. when a speech is interrupted by a question of another legislator, or by the chair. Concerns about disturbing noises such as applause are mitigated by the fact that `Praat`, by default, implements a "voicing threshold" to discriminate between voiced and unvoiced segments of an audio file. The software then only uses voiced segments to calculate the mean fundamental frequency.[1] Still, there is no guarantee that there are no remaining disturbances that add noise to the measurement, especially if the disturbances stem from other speakers. The analysis thus relies on the assumption that the measurement error in vocal pitch introduced by such disturbances does not confound the estimated relationship between

[1] I rely on `Praat`'s default voicing threshold (0.45).

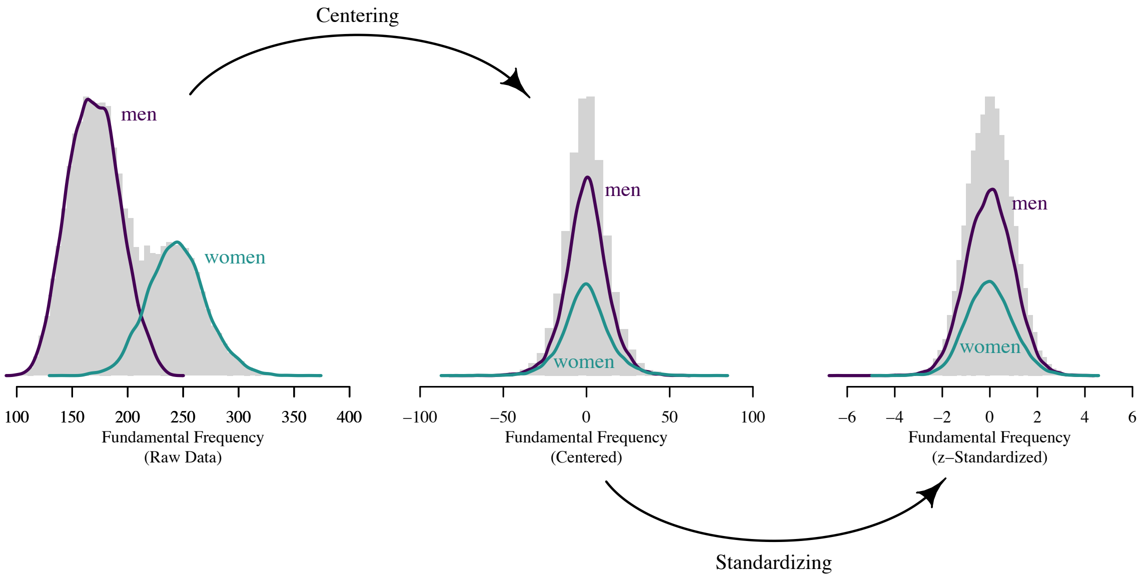

Figure A3: Visualization of standardization procedure of the mean fundamental frequency (average vocal pitch). The left panel shows the distribution of average vocal pitch values for all speeches in the data set. Density curves additionally highlight how this distribution is composed of speeches by men and women. In the middle panel, the distribution is centered around the speaker's average levels of vocal pitch. In the right panel, the distribution is rescaled to reflect standard deviations from the speaker's average level of vocal pitch.

the topic of a speech, gender of the speaker, and vocal pitch, *after* being normalized by speakers.

C Matching of Audio and Text Data

For the analysis, it is necessary to create a data set that combines audio and text data within one data set. The audio data stems from the online video archive of the German Bundestag (see Appendix B). The unit of observation of this data set is individual speeches. In other words, one entry in this data set refers to one recording of a single speech. The text data stems from two sources: For the period between 2011 and 2018, I rely on the ParlsSpeech V2 data set (Rauh and Schwalbach, 2020). As for the audio data, the unit of observation of this data set is individual speeches. For the period between 2018 and 2020, I rely on transcripts based on plenary protocols that can be downloaded from the Bundestag homepage. These transcripts are available in `.xml` format with individual speeches as units of observation.

The goal of the audio-text-matching is to find the correct transcript for each speech audio recording. In the following, I explain all steps undertaken to achieve this task.

Step 1: Create a common legislator identifier in all data sets.

In the first step, I made sure that both the audio and text data sets contain a common identifier variable for each person who appears as a speaker in any of the data sets. For the vast majority of speaker names, this was possible based on the name itself. Whenever this was not possible based on speaker names, e.g. due to inconsistent spelling of a name across data sets, the task was handled manually.

Step 2: Create a common agenda item identifier in text and audio data.

Next, I made sure that both the audio and text data sets contain a common identifier variable for the agenda item that a speech belongs to. Conveniently, all three data sets contain an agenda item identifier in their original format. However, the annotation logic of agenda items varies between the data sets, preventing a simple matching procedure. To solve this, I manipulated the agenda item identifier in the two text data sets so that they match the agenda item identifier in the audio data set. This was done using regular expressions and manual manipulation. At the end of this process, all speeches in both the text and video data contained a distinct and harmonized identifier of the agenda item they belong. Thus, the remaining task was to match audio and text data *within* agenda items.

Step 3: Match audio and text data of speeches within agenda items.

Prior to the matching, I identify speeches by the chair and exclude them from both audio and text data sets. I also exlude all speeches that belong to a question hour (*Fragestunde*) or questions to the government (*Befragung der Bundesregierung*)[2] agenda items. I implemented a series of matching algorithms to match audio and text data. In the following, I describe each of those algorithms. I also report how many audio files were matched with text data based on each algorithm.

Algorithm 1:

```
Loop over all agenda items:
  → Check whether audio and text data within the agenda item contain exactly the
    same sequence of speakers.
      → If TRUE: matching is straightforward.
      → If TRUE: move to next agenda item.
```

Example: In the following example, the exact same sequence of speakers are present in the audio and text data (within one agenda item). They can be matched unambiguously.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker B	T2. Speaker B
A3. Speaker C	T3. Speaker C
A4. Speaker D	T4. Speaker D

Success:
Parlspeech: 4,160 / 26,231 (+4,160)
Transcripts: 8,404 / 8,565 (+8,404)

[2] 303 audio recordings belonging to *Befragung der Bundesregierung* were excluded after matching them to text data.

Algorithm 2:

Loop over all agenda items with zero matched speeches:

→ Check whether the sequence of speakers is equal in both data sets, but the number of speeches differs (e.g. more speeches in text data).

→ If `TRUE`: concatenate speech text of two or more consecutive speeches by one legislator to one speech.

→ If `FALSE`: move to next agenda item.

Example: In the following example, the text of the two speeches by Speaker B (T1 and T2) would be concatenated into one speech and matched with A2 in the audio data set.

This pattern could occur if Speaker B was shortly interrupted by the chair, e.g. because they were going over time. This could trigger a "new speech" in the text data while the audio recording continues.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker B	T2. Speaker B
	T3. Speaker B
A3. Speaker C	T4. Speaker C
A4. Speaker D	T5. Speaker D

Success:

Parlspeech: 11,735 / 26,231 (+7,575)

Transcripts: 8,404 / 8,565 (+0)

Algorithm 3:

Loop over all agenda items with zero matched speeches:

→ Check whether there is a speaker in text data who does not appear in audio data.

→ If `TRUE`: Delete speeches from the speaker who appears in text data but not in video data and apply matching algorithm 1.

→ If `FALSE`: move to next agenda item.

Example: In the following example, the speech by Speaker C (T3) does not appear in the audio data and is deleted. After deletion of T3, the sequences are equal to each other and can be matched using algorithm 1.

This pattern could occur, if the statement by Speaker C was classified as procedural and thus excluded in the audio data set (e.g. a statement by the chair), but not in the text data set.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker B	T2. Speaker B
	T3. Speaker C
A3. Speaker D	T4. Speaker D
A4. Speaker E	T5. Speaker E

Success:

Parlspeech: 11,854 / 26,231 (+119)

Transcripts: 8,456 / 8,565 (+52)

Algorithm 4:

Loop over all agenda items with zero matched speeches:

→ Check whether there is a speaker in text data who does not appear in audio data.

→ If `TRUE`: Delete speeches from the speaker who appears in text data but not in video data and apply matching algorithm 2.

→ If `FALSE`: move to next agenda item.

Example: In the following example, the speech by Speaker C (T4) does not appear in the audio data and is deleted. After deletion, the sequences match and can be matched using algorithm 2.

This pattern could occur when Speaker B was interrupted twice. Once by the chair (between T2 and T3) and once by Speaker C, e.g. Speaker C asked a question. Both interruptions could trigger new speeches in the text data, while the audio recording would continue.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker B	T2. Speaker B
	T3. Speaker B
	T4. Speaker C
	T3. Speaker B
A3. Speaker D	T5. Speaker D
A4. Speaker E	T6. Speaker E

Success:

Parlspeech: 15,483 / 26,231 (+3,629)

Transcripts: 8,456 / 8,565 (+0)

Algorithm 5:

```
Loop over all agenda items with zero matched speeches:
  → Loop over all unique speakers who appear in the audio data.
      → Check whether the speaker occurs exactly once in the audio data.
          → If TRUE: Check whether the speaker occurs exactly once in the text data.
              → If TRUE: Match the speech.
              → If FALSE: Move to next speaker.
          → If FALSE: Move to next speaker.
```

Example: In the following example, even though speeches by speaker A and B are difficult to match unambiguously, speaker C occurs exactly once in both sequences and their speech can be matched.

This pattern could occur if Speaker A gave a speech (T1) and later interrupted Speaker B, for example by asking a question (T3).

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker B	T2. Speaker B
	T3. Speaker A
	T4. Speaker B
A3. Speaker C	T5. Speaker C

Success:

Parlspeech: 21,220 / 26,231 (+5,737)

Transcripts: 8,543 / 8,565 (+87)

Algorithm 6:

Loop over all unmatched speeches:

→ Check whether the speaker occurs as often in the audio and text data within the agenda item.

→ → If `TRUE`: Match the speeches in their order.
→ If `FALSE`: Move to next speech.

Example: In the following instance, speaker C occurs twice in both the text and audio data. The algorithm matches A2 with T4 and A4 with T6. For this particular agenda item, A2 and T2 would have been matched in the previous round by algorithm 5.

This pattern could occur of both Speaker A and Speaker C delivered two speeches, but Speaker A was interrupted by Speaker B in their first speech (T2), e.g. because Speaker B asked a question.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker C	T2. Speaker B
	T3. Speaker A
	T4. Speaker C
A3. Speaker A	T5. Speaker A
A4. Speaker C	T6. Speaker C

Success:

Parlspeech: 21,535 / 26,231 (+315)

Transcripts: 8,549 / 8,565 (+6)

Algorithm 7:

Loop over all agenda items with zero matched speeches:

→ Check whether there are sequences of one or more consecutive speeches by one legislator in the text data.

  → If `TRUE`: Combine the consecutive speeches into one speech and apply algorithm 6.

  → If `FALSE`: Move to next agenda item.

Example: In the following example, the algorithm concatenates T1+T2, and T4+T5. Next, it matches the speeches by Speaker A, (T1+T2) with A1, and the speeches by Speaker B, (T4+T5) with A2.

This pattern could occur of both Speaker A and Speaker B delivered a speech, and were both interrupted by the chair, e.g. because they were going over time. Both interruptions could trigger a new speech in the text but not in the audio data.

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
	T2. Speaker A
	T3. Speaker B
A2. Speaker C	T4. Speaker C
A3. Speaker B	T5. Speaker C
	T6. Speaker B

Success:

Parlspeech: 22,875 / 26,231 (+1,340)

Transcripts: 8,549 / 8,565 (+0)

Algorithm 8:

Loop over all agenda items with zero matched speeches:

→ Loop over all unique speakers in audio data and check whether they occur exactly once within the agenda item.

- → If `TRUE`: Search for all speeches by the speaker in text data. Apply the dictionary to all speeches in the text data and store results. Concatenate all speeches irrespective of their position within the agenda item and apply the dictionary again. If all dictionary classifications are equal, use the concatenated text. Code manually if the dictionary classifications differ between the original texts and the concatenated text.
- → If `FALSE`: Move to next agenda item.

Example: In the following instance, it is not possible to unambiguously match the speech by Speaker A (A1) with text data. In this case, the algorithm starts by concatenating all text data that belongs to Speaker A within the agenda item (T1, T3, T5). The concern with this procedure is that the concatenated text includes more than what Speaker A says during the audio recording A1, and that this may introduce error to the dictionary classification in the next analysis step. To make sure that this is not the case, the algorithm applies the dictionary to classify the concatenated text and to T1, T3, and T5 separately. This yields four dictionary classifications. The algorithm then checks whether all classifications are equal, or whether they diverge. The algorithm matches A1 with the concatenated text of T1, T3, and T5 if all four classifications are equal, i.e. if the concatenation does not make a difference for the dictionary classification. If the classifications diverge, a human coder watches the recording of A1 and manually matches A1 with the correct text data.

This pattern could occur if Speaker A was interrupted by Speaker B in their speech, e.g. because Speaker B was asking a question (T2). Also, Speaker C asked a question during the speech by speaker C (T5).

Speaker Sequence in Audio Data	Speaker Sequence in Text Data
A1. Speaker A	T1. Speaker A
A2. Speaker C	T2. Speaker B
	T3. Speaker A
	T4. Speaker C
	T5. Speaker A
	T6. Speaker C

Success:

Parlspeech: 25,024 / 26,231 (+2,149)

332 of the 2,149 speeches were coded manually because the dictionary classification of the concatenated version differed from the classification of its components.

Transcripts: 8,549 / 8,565 (+0)

Algorithm 9: Manual Matching

In the Parlspeech period, 1,207 audio recordings could not be matched based on the prior eight algorithms. Of those unmatched audio recordings...

→ 554 speeches were matched manually.

→ 310 belonged to question hours and were excluded.

→ 251 speeches were procedural (mostly prior unidentified speeches by the chair) and excluded.

→ 92 speeches remain unmatched.

In the Transcripts period, sixteen audio recordings could not be matched based on the eight algorithms. Of those unmatched audio recordings...

→ eight speeches were matched manually.

→ three were manually classified as procedural statements by the chair and excluded.

→ five speeches remain unmatched.

Validation of the Audio-Text-Matching Process

I validated the audio-text-matching process based on a random sample of 100 speeches. For each speech, I checked whether the date and speaker are correct and whether the audio recording is matched to the correct text of the speech. All 100 speeches were matched correctly. In one instance, the text exceeds what the speaker says in the audio recording of their speech, but without affecting the dictionary classification. The respective speech was matched based on algorithm 8, so the minor mismatch is within the logic of the matching procedure. There was an additional speech where the speech text included information on the following vote. This, too, did not affect the dictionary classification. Overall, I conclude that the procedure matches audio recordings and speech text reliably.

D Effective Sample Size and Missing Values

As described above, a total of 34,135 audio recordings of speeches in the German Bundestag were matched with text data. Because these audio recordings were matched successfully with the text of the speech, there are no missing values regarding the topic of the speech. There are also no missing values regarding the gender of the speaker. Still, the statistical model in table 1 (dependent variable: women mentioned) leverages 33,514 observations. Thus, there is a discrepancy of 621 observations. Here, I explain the source of this discrepancy.

- 303 matched speeches were excluded because they were belonged to “questions to the government"-agenda items (*Befragung der Bundesregierung*).
- 260 speeches come from speakers who are not members of the German Bundestag.
- 58 speeches with less than 50 words.

The statistical model in figure 3 (dependent variable: vocal pitch) is based on 33,489 observations, 25 less than the model in table 1. Six observations are missing because they had corrupted audio files. 21 speeches come from legislators who delivered only one speech and therefore had no variation in the dependent variable.

E Translation of the Pearson and Dancey (2011)-Dictionary

original dictionary	translated dictionary
woman	frau*
women	frauen*
woman's	
women's	
girl	mädchen*
girl's	
girls	
girls'	
female	weiblich*
female's	
females	
females'	
servicewoman*	schwester*
servicewoman's	
servicewomen*	
servicewomen's	

Table A1: Translation of the dictionary by Pearson and Dancey (2011). Asterisk indicates that word extensions are included as well.

Table A1 shows the English language dictionary by Pearson and Dancey (2011) side by side with the translated version. The translated version's main concern is the German word for woman. This concern stems from the fact that the German translation, *Frau*, is used to address women in German (e.g. *"Sehr geehrte Frau Merkel"*). This habitual and frequent use of the word *Frau* implies that a legislator addresses a specific woman (e.g., a fellow women legislator) but does not imply that the speech addresses women-related issues. To address this concern, I clean the texts with respect to this use of the term *Frau* as a form of address before applying the dictionary. In doing so, I avoid the dictionary falsely classifying speeches as women-related only because the speaker used the term *Frau* as a form of address in their speech without speaking about women in general.

E.1 Text cleaning

I use a regular expression (regex), `Frau\s[A-Z].+?\w*`, to detect and delete the word Frau whenever it is used as a form of address. Put simply, this regex searches for occurrences of the term `Frau`, followed by a whitespace and a second word starting with an uppercase letter. In German, only nouns and proper nouns begin with uppercase letters.[3] At the same time, German grammar does not provide cases where two nouns directly follow each other. Thus, if the word `Frau` is followed by another word that starts with an uppercase letter, the second word is almost certainly a name. In that case, the term `Frau` is used as a form of address.

Still, the regex is not error-free. We can think of the regex as a classifier that classifies each occurrence of the term `Frau` as being used as a form of address or not. There are two types of error: False negatives and false positives. False negatives describe cases where the regex does not classify an occurrence of the term `Frau` as a form of address even though it was used as such. False positives describe cases where the regex classifies an occurrence of the term `Frau` as a form of address even though it is not used as such.

To address false negatives, I manually check all occurrences of the word `Frau` that were not classified as a form of address but are decisive for the dictionary classification. Those are speeches where the term `Frau` is the only dictionary term in the speech. The term `Frau` and its classification as form of address is crucial in such speeches: The dictionary classifies the speech as women-related *only* if the term `Frau` is not classified as a form of addresses within the particular speech. By checking all occurrences of the term `Frau` where this is the case, I fully eliminate dictionary misclassification based on this type of error. In total, I checked 1,436 occurrences of the term `Frau`. 657 of these occurrences (45,8%) were identified as false negatives. For example, the regex misclassifies `Frau von der Leyen` (Minister of Interior 2009–2013 and Minister of Defence 2013–2019) because her name starts with a lowercase letter. Another reason for false negatives are occurrences

[3] Words at the beginning of a sentence start with an uppercase letter, too, but this is irrelevant to the issue at hand.

of the term Frau, where Frau was part of another word that is unrelated to women, e.g., Fraunhofer Institut, a research institute in Germany.

For each unique expression that led to a false negative classification, I created another regex and applied it to the text before the dictionary application. This led to the following list of regular expressions: Fraunhofer (research institute), Frau Özo (Frau Özoğus, missed due to encoding issue), Frau von der Leyen, Frau von Storch, Frauke Petry, Frau iffey (Frau Giffey, spelling error in text), Frau Özu (Frau Özoğus, spelling error in text), Frau de Ridder, Frau de la Durantaye, FrauVerlinden (Frau Verlinden, spelling error in text), Frau rugger (Frau Rugger, spelling error in text), Frauke Brosius-Gersdorf, Frau al-Basri, Frau, Hagedorn, Frau von Cramon, Frau Öney (missed due to encoding issue), \sFraud\s (e.g. "Fraud and Corruption Network"), Community Fraud, Frau de Sutter, Anti.Fraud.Behörde.

Addressing false positives is more difficult. A false positive is an occurrence of the term Frau that was classified as a form of address (and thus excluded from the text) even though it was not used as a form of address. It is not possible to manually control all occurrences of the term Frau that were filtered out by the regex (>35.000 instances). To assess whether false positives constitute a problem, I inspect a random sample of 500 positives and manually code whether they are true or false positives (i.e., whether the term Frau was indeed used as a form of address). None of the 500 instances were falsely classified as form of address. Thus, I conclude that the regex is not prone to false positive classification.

I apply the dictionary to the speech corpus after cleaning the text with respect to the term Frau as a form of address as described above. This reveals that legislators in the German Bundestag mention at least one of the dictionary terms in 14.7% of their speeches. This is reasonably close to the number reported by Dietrich, Hayes and O'Brien (2019), who find that members of the US House of Representatives do so in 10.7% of their speeches. In the next section, I report the results of a more rigorous validation test.

E.2 Dictionary Validation

I assess whether the translated dictionary for Bundestag speeches performs similarly to the original dictionary for US House speeches. To that end, a student assistant was instructed to read 500 randomly selected speeches of the Bundestag corpus (2011–2020) and indicate for each of them whether the speaker (explicitly) mentioned women in their speech or not. The same student assistant also read 500 randomly selected speeches delivered in the US House of Representatives with the same labeling instructions. Note that the speech corpus used by Dietrich, Hayes and O'Brien (2019) is not part of their replication data. For that reason, I rely on the `Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts`-Dataset by Gentzkow, Shapiro and Taddy (2018). This data set contains the processed text of speeches in the US Congress. The random sample of 500 speeches was drawn from a subset of this data set, including speeches delivered in the US House of Representatives between January 6, 2009, and August 4, 2014. This matches the time frame analyzed by Dietrich, Hayes and O'Brien (2019).

Table A2 shows the result of the validation exercise. The hand-coded labels are consistent with the dictionary labels in 85.8% of the US House speeches and 90.6% of the Bundestag speeches. However, this accuracy score needs to be put into context as the categories are not balanced: Only 8.4% of the US House and 15.1% of the Bundestag speeches mention women, according to the research assistant labels. Therefore, the table also presents sensitivity (true-positive rate) and specificity (true-negative rate) scores and F1 scores for both sets of speeches. The US dictionary correctly identified speeches about women in 69.0% (SE = 7.1 percentage points) of the cases. This score is almost identical to the German dictionary, which correctly identified speeches about women in 68.4% (SE = 5.6 percentage points) of the cases. The US dictionary shows a slightly higher tendency to incorrectly classify speeches not about women as speeches about women (i.e., a higher false-negative rate). This is reflected in a lower specificity score of the English language dictionary (0.873, SE = 0.016) compared to the German language dictionary (0.946, $SE = 0.011$).

	US House	Bundestag
Accuracy	0.858	0.906
	(0.016)	(0.013)
Sensitivity	0.690	0.684
	(0.071)	(0.053)
Specificity	0.873	0.946
	(0.016)	(0.011)
F1	0.450	0.689

Table A2: Validation of dictionary classifications based on 500 hand-coded speeches of speeches in the US House of Representatives and the German Bundestag, respectively.

The validation exercise also demonstrates the necessity of cleaning the text with respect to the term "Frau" (*eng.:* woman) whenever it is used as a form of address by the speaker. Applying the dictionary to the German text before removing "Frau" as form of address substantially changes the performance scores. The rationale behind removing the term "Frau" when used as a form of address primarily relates to concerns about the false-positive rate, i.e., the dictionary classifies speeches to be about women, *only* because the speaker addresses another woman in their speech (e.g., "Sehr geehrte Frau Präsidentin", *eng*: "Dear Mrs. President"). Indeed, without previous text cleaning, a high number of the speeches that are *not* about women would be classified as being about women. This would disproportionately increase the false positive rate and lead to a much worse overall outcome. More precisely, if I did not exclude the word "Frau" when used to refer to other women, the overall accuracy of the German language dictionary would decrease to 0.486 ($SE = 0.022$), and the specificity score would drop to 0.423 ($SE = 0.024$).

Overall, the validation exercise shows only minor differences in the performance of the translated dictionary compared to the original dictionary when benchmarked against human classifications. I consider these differences small enough so that the measurement of whether a speaker mentions women in their speech is comparable in both studies.

F Statistical Power Analysis

The sample with speeches in the Bundestag comprises 32,897 speeches. With that, it is less than half as large as the sample used by Dietrich, Hayes and O'Brien (2019), who base their analysis on 71,198 congressional speeches in the US House of Representatives. Here, I approximate the statistical power of the replication study to assess whether this poses the risk of too little statistical power. This may seem overly cautious given the still large number of observations. Three reasons speak against this assessment: First, the differences in vocal pitch between men and women when talking about women found by Dietrich, Hayes and O'Brien (2019) are subtle. When talking about women, the average difference in z-standardized vocal pitch between men and women amounts to 0.094 standard deviations. Detecting such subtle differences may require large amounts of data. Second, previous research found that quantitative political science is notoriously underpowered and, third, that even political methodologists tend to overestimate the statistical power of quantitative studies in their field (Arel-Bundock et al., 2022).

To approximate the statistical power of the replication study, I follow an approach inspired by Blair et al. (2019). Specifically, I use a simulation approach to approximate the statistical power to detect statistically significant estimates of the two primary quantities of interest at $\alpha = 0.05$. That is, first, a statistically significant difference in the average vocal pitch of women representatives when they use one of the "women"-dictionary terms in their speech versus the average vocal pitch of women representatives when they do not use one of the dictionary terms. This refers to the marginal effect of "women mentioned" conditional on being a woman as presented in the right panel of figure 3. Second, the differences in average vocal pitch between men and women representatives when referencing women in a plenary speech (Blair et al., 2019). This refers to the interaction term in the table in the left panel of figure 3. I am interested in the statistical power to detect statistically significant estimates τ of the respective differences based on a 95% confidence level ($H_0 : \tau = 0$, $H_a : \tau \neq 0$), given that the true coefficients are at least as large as those found based on data from the US, reported in the left column of the table in figure 3.

To estimate power, I conduct the following simulation:

1. I start with observed data X. The unit of analysis are speeches by legislators in the German Bundestag (or the US House of Representatives). X contains information on whether at least one of the Pearson and Dancey (2011) dictionary terms was used in the speech, and the gender of the speaking legislator for all speeches included in the compiled data set used for the analysis in the paper. Because the statistical model includes legislator fixed effects, I recompute the variable "women mentioned" so that it represents the deviation from individual level average values.

2. I simulate y as $y \sim N(X\beta, \sigma^2)$, where β corresponds to the reported point estimates in the left column of the table in figure 3, σ^2 to the residual variance of the respective model. X is the matrix of the covariates derived in the previous step.

3. I regress y on "Women" mentioned, female speaker, and an interaction term of both. Based on the estimation results, I calculate whether (a) the marginal effect of "women mentioned" conditional on being a female speaker and (b) the interaction term between "women mentioned" and "female speaker" are significantly different from zero at $\alpha = .05$.

4. I repeat step 1–3 $s = 5000$ times. For each quantity respectively, the power estimate is given by the number of significant coefficients in step 3 divided by s.

Table A3 reports the results. To establish a baseline, I also report the results of a post hoc power approximation of the original study in the US House. The results show that the replication is perfectly powered to detect whether women speak, on average, at a higher vocal pitch when they reference women versus when they do not reference women. The replication study has a power of 99.6% to detect differences that are at least as large as the ones found in the US House of Representatives. Thus, if women representatives speak at a higher vocal pitch when they reference women, the replication study almost certainly reveals them.

		Power to reject $H_0 : \tau \neq 0$	
		US House ($N = 71,198$)	Bundestag ($N = 32,897$)
(1)	$\mathbb{E}[Y_{is} \mid W_{is} = 1, F_i = 1] - \mathbb{E}[Y_{is} \mid W_{is} = 0, F_i = 1]$	0.998 (0.001)	0.996 (0.001)
(2)	$(\mathbb{E}[Y_{is} \mid W_{is} = 1, F_i = 1] - \mathbb{E}[Y_{is} \mid W_{is} = 0, F_i = 1])-$ $(\mathbb{E}[Y_{is} \mid W_{is} = 1, F_i = 0] - \mathbb{E}[Y_{is} \mid W_{is} = 0, F_i = 0])$	0.917 (0.004)	0.794 (0.006)

Table A3: Simulation-based statistical power approximation with standard errors in parentheses for the two main quantities of interest ($\alpha = .05$). W_{is} denotes whether legislator i mentioned women in speech s, F_i denotes whether legislator i identifies as female. Quantity (1) represents the average difference in women's vocal pitch when they mention women versus their average vocal pitch when do not mention women. Here, the replication is well-powered. Quantity (2) represents the average difference in how vocal pitch changes for women when they mention women, versus how it changes for men when they mention women. Here, the statistical power for the Bundestag drops to 79.4%.

Second, I consider whether the replication study is well-powered to detect a statistically significant interaction coefficient between "women mentioned" and "female". This is an interesting estimand because rejecting the null hypothesis of no difference means that women adjust their vocal pitch differently when mentioning women than men. Conversely, the hypothesis cannot be rejected if both men and women elevate their vocal pitch in the same way when mentioning women. Here, the post hoc power approximation of the US study reveals a statistical power of 91.7%. The statistical power of the replication study drops to 79.4%. If the true size of the interaction term is as large as the one reported by Dietrich, Hayes and O'Brien (2019), this means that the replication study's probability of finding a statistically significant interaction term is 79.4%. This also means that there is a fair chance of not finding a statistically significant interaction term, even if there is a difference between men and women. This must be kept in mind in such a scenario.

The result of the power approximation also signals that splitting the data further, e.g. by including another interaction term in the model, very likely leads to underpowered estimates. For example, it might be of interest to investigate differences between members of the Bundestag who were elected in their local district and members of the Bundestag who gained their mandate through the party list. The results of the power approximation strongly advise against using 95% significance tests to investigate such differences. When estimators are underpowered, statistically non-significant differences provide little to no evidence in support of the Null hypothesis. At the same time, statistically significant estimates are likely to be over- or underestimated.

References

Arel-Bundock, Vincent, Ryan Briggs, Hristos Doucouliagos, Marco Mendoza Aviña and T.D. Stanley. 2022. "Quantitative Political Science Research Is Greatly Underpowered." OSF Preprints. July 5. `https://doi.org/10.31219/osf.io/7vy2f`.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113(3):838–859.

Boersma, Paul and David Weenik. 2021. "Praat: doing phonetics by computer.". [computer program], version 6.1.51. Retrieved from: `http://www.praat.org/`.

Dietrich, Bryce J., Matthew Hayes and Diana Z. O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4):941–62.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2018. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts.". data retrieved via Palo Alto, CA: Stanford Libraries [distributor] at `https://data.stanford.edu/congress_text`.

Inter-Parliamentary Union. 2022. "Historical Data on Women in National Parliaments.". **URL:** *https://data.ipu.org/historical-women*

Pearson, Kathryn and Logan Dancey. 2011. "Speaking for the Underrepresented in the House of Representatives: Voicing Women's Interests in a Partisan Era." *Politics & Gender* 7(4):493–519.

Rauh, Christian and Jan Schwalbach. 2020. "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.".
URL: *https://doi.org/10.7910/DVN/L4OAKN*