Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts

Matthew Gentzkow, *Stanford University and NBER*[*]

Jesse M. Shapiro, *Brown University and NBER*

Matt Taddy, *Microsoft and Chicago Booth*

January 16, 2018

1 Organization

This corpus contains three groups of files: (i) speeches, counts of stemmed two-word phrases (bigrams), and speech-level metadata from each session of Congress; (ii) vocabulary from all sessions of Congress; and (iii) validation materials. Unless otherwise noted, each file has a header as its first row and is pipe-delimited, not wrapped by quotes, and saved without row numbers.

1.1 Speeches, Counts, and Metadata

There are five types of files in this group. Each file is produced for every session of Congress and edition of the *Congressional Record* that is available from HeinOnline. Table 1 explains their purposes. Files from the bound edition are stored in hein-bound, files from the daily edition are stored in hein-daily, and ### is a stand-in for each file's three-digit session number.

1.2 Vocabulary

The bigrams from all sessions are collected and labeled to produce the overall vocabulary used in Gentzkow et al. (2017). Vocabulary files are stored in vocabulary and organized according to the scheme in Table 2.

[*] E-mail: gentzkow@stanford.edu, jesse_shapiro_1@brown.edu, matt.taddy@chicagobooth.edu

Table 1: Format of Speech Files

File	Content	Key	Format
byparty_2gram_###.txt	bigram counts by party	party, phrase	2-column
byspeaker_2gram_###.txt	bigram counts by congressperson	speakerid, phrase	3-column
descr_###.txt	speech metadata	speech_id	14-column
speeches_###.txt	full-text speeches	speech_id	2-column
###_SpeakerMap.txt	speaker metadata	speech_id	10-column

Table 2: Format of Vocabulary Files

File	Content	Key	Format
master_list.txt	all bigrams and labels	phrase	2-column
procedural.txt	procedural bigrams and labels	phrase	2-column
vocab.txt	valid bigrams	———	1-column, no header, no delimiter

1.3 Validation

Validation materials are provided in two groups: (i) input and output from a manual audit and (ii) the match rates between speeches and speakers.

1.3.1 Manual Audit

The manual audit involves two checks made separately on the bound and daily editions of the *Congressional Record*: manual parsing and manual speaker matching. Information on the parsing and matching is in Tables 3 and 4. The date is in month, day, year format as MMDDYYYY; ### is a stand-in for a three-digit session number; and * is a greedy wildcard character.

1.3.2 Match Statistics

The statistics for matching speeches to members of Congress are documented for each session of Congress and edition of the *Congressional Record*. Table 5 gives the file structure. The statistics are the number of speeches matched with a congressperson, the number of unmatched speeches, and the match rate.

1.4 File sizes

Files are zipped for faster downloading. Table 6 gives the sizes of the zipped archives and the sizes and counts of their unzipped contents.

Table 3: Format of Files for Manual Parsing and Audit

File	Content	Key	Format
descr_MMDDYYYY.txt	speech metadata, automated	speech_id or line_start or line_end	14-column
speeches_MMDDYYYY.txt	full-text speeches, automated	speech_id or line_start or line_end	2-column
manual/*/MMDDYYYY.txt	manual parsing	LINE_START or LINE_END	5-column
raw/*/MMDDYYYY.txt	lines of text to manually parse	———	raw text, no header, no delimiter
comparison_master.txt	overall summary	———	human readable, no header, no delimiter
comparison_###.txt	summary for a session of Congress	———	human readable, no header, no delimiter
errors_###.txt	discrepancies between parsings	file, line_start	6-column
audit_selections-hein-X.txt	day level files and lines to audit	filename	4-column

Table 4: Format of Files for Manual Matching of Speakers to Congresspeople

File	Content	Key	Format
hein-X_errors.txt	All speech_id from manual matching that could not be matched in any ###_SpeakerMap.txt.	speech_id	15-column, row numbers in first column, empty for daily edition

Table 5: Format of Files with Statistics on Matching of Speeches to Congresspeople

File	Content	Key	Format
hein-X.txt	statistics on matching speeches to congresspeople	cong	4-column

Table 6: Sizes of Archives

Archive	Zipped Size	Unzipped Size	Unzipped Count
audit.zip	2.3 Mb	5.7 Mb	385 files
hein-bound.zip	7.6 Gb	31.2 Gb	345 files
hein-daily.zip	2.8 Gb	11.4 Gb	90 files
speakermap_stats.zip	3.2 Kb	2.7 Kb	2 files
vocabulary.zip	41.8 Mb	205.5 Mb	3 files

2 Data Processing

2.1 Raw Data

HeinOnline scans print volumes of the bound and daily editions of the *Congressional Record.* The scans are converted to plain text files through optical character recognition (OCR). Each file corresponds to a single page of a print volume, with metadata stored in a separate file for each session of Congress and edition of the *Congressional Record.* The metadata is used to put files in the same order as pages in the print volumes and match files with days of speech they record. All files from the same day are then concatenated in order, stamped with their filenames, and saved as text files with names in month-day-year format. Raw text is assigned to a session of Congress by aligning these day-level text files with the dates of each session.

2.2 Speeches

An automated script parses the day-level text files line-by-line. The parser encodes the structure of the day-level text files, which is itself a product of the organization of the print volumes, the OCR, and the concatenation. This encoding enables the parser to avoid non-speech text and sort the remaining text into distinct speeches.[1]

All speeches in the *Congressional Record* begin on a new line with a standardized demarcation of the speaker: (i) a titled speaker's position (e.g., The SPEAKER pro tempore) or (ii) a standard speaker's title and last name in uppercase (e.g., Mr. ALLEN), sometimes including the speaker's state and uppercase first name as well (e.g., Mr. JOHN ALLEN of Illinois). The parser recognizes the start of a speech when it encounters a line of text that begins with this demarcation and is not an obvious reference to a speaker made during another's speech. From the start of a speech, the parser assigns all lines of text to that speech until it reaches the speech's end. The parser recognizes the end of a speech when it reaches the beginning of a new speech, the end of a section of debate or day of Congress, the reading of a bill or other document, or a vote.

The text comprising the speaker and speech is cleaned and then processed. Cleaning proceeds in two steps. First, the text of the speaker demarcation is separated from the text of the speech. Second, the text of the speech is processed by (i) removing non-speech text, (ii) removing apostrophes and replacing commas and semicolons with periods, (iii) replacing repeated whitespace characters with a single space, (iv) removing punctuation—hyphens, periods, and asterisks—that separate the speaker's demarcation from the speech, and (v) removing whitespace leading and trailing the speech.

[1] Non-speech text comprises of page headers and footers, section titles, parenthetical insertions, votes, and administrative time allotments.

Table 7: Sources of Speech Metadata

Attribute	Location
chamber	page headings
date	file metadata described in Section 2.1
file	name of day-level file in which a speech is found
line start	line number on which a speech starting condition from Section 2.2 occurs
line end	one less than the line number on which a speech ending condition from Section 2.2 occurs
order within day-level file	running count of parsed speeches

Speech processing comprises seven steps. First, the number of characters and space-delimited words are computed. Second, the speech is coerced to lowercase. Third, the speech is broken into separate words, treating all non-alphanumeric characters as delimiters. Fourth, general English-language stopwords are removed.[2] Fifth, remaining words are reduced to their stems using the Porter2 (English) stemming algorithm.[3] Sixth, the stemmed words are converted to bigrams following their order in the speech. Seventh, the bigrams of the speech are converted into counts of bigrams, which undoes the ordering.

2.3 Speech Metadata

The parser records speech metadata from the sources described in Table 7.

2.4 Speaker Metadata

Metadata on speakers is initially drawn from speaker-demarcation conventions. Speeches are then paired with members of Congress by matching these initial metadata against complete congressperson characteristics from a historical source.[4] Matches are performed separately for each session of Congress and edition of the *Congressional Record*.

The matching procedure creates sets of matchable characteristics for each congressperson. First, an initial set of characteristics—chamber, gender, first name, last name, and state—is taken from the historical source for each congressperson. Second, the set of all subsets (the power set) of these initial characteristics is

[2] The set of stopwords is defined by a list obtained from http://snowball.tartarus.org/algorithms/english/stop.txt on November 11, 2010.

[3] The Python implementation of the algorithm is taken from https://pypi.python.org/pypi/PyStemmer/1.3.0.

[4] The source is the congress-legislators GitHub repository https://github.com/unitedstates/congress-legislators/tree/1473ea983d5538c25f5d315626445ab038d8141b. The source records all the information available from the speaker demarcation as well as each congressperson's congressional district, party, and voting status at the end of each session of Congress. Separate entries are recorded when a congressperson serves in both the House of Representatives and the Senate during a single session of Congress. Entries from congress-legislators can be matched to this corpus by their natural keys: first and last name, gender, party, chamber of Congress, session of Congress, state, and district.

computed for each congressperson. Third, subsets are removed if they are not unique among congresspeople in a single session.

The partial metadata recovered from the *Congressional Record* is then checked for a unique match against the sets of matchable characteristics for each session of Congress. Speeches with unique matches are assigned to the associated congressperson. Speeches without a unique match, made by titled speakers, or found in the Extensions of Remarks are dropped.

Unique matches are determined according to a fuzzy matching algorithm. Exact matches are required for a speaker's chamber and gender. A speaker's first name may match exactly or differ by a single simple edit.[5] The same rule applies to a speaker's last name, but manual corrections are made to misspellings that are obvious but difficult to express as simple edits (e.g., when a word is excluded from a multi-word last name). A speaker's state may differ by up to two simple edits.

2.5 Bigram Counts

The counts of bigrams from each speech are aggregated by congressperson, separately for each session of Congress and edition of the *Congressional Record*. The counts are further aggregated by political party: Democratic, Republican, and independent or third-party congresspeople. Both aggregations exclude bigrams spoken by non-voting delegates.

2.6 Vocabulary

This processing step produces the final vocabulary used by Gentzkow et al. (2017) and described in that paper's Online Appendix. See Appendix A for details.

3 Manual Audit

Two manual audits check the accuracy of the algorithms for parsing the *Congressional Record* and matching speeches to speakers.

3.1 Parsing

The manual parsing audits text selected by a multistage sampling procedure that chooses 2539 (319) speeches from the bound (daily) edition. The procedure randomly selects a two-year period among the first 10 (5) periods of the bound (daily) edition and randomly samples four days for inspection. It then samples four

[5] Simple edits include the insertion of an incorrect letter, deletion of a correct letter, replacement of a correct letter by an incorrect letter, or transposition of two adjacent letters. Simple edits that do not produce a unique match are excluded as well.

days from every subsequent tenth (fifth) period, until no more remain. On each sampled day, it selects 1,000 contiguous lines of text by randomly choosing a start line and the following 999 lines of text. Start lines are required to be outside the last 1,000 lines of text if there are more than 1,000 lines in a day, and all the lines in a day are selected if there are 1,000 or fewer.

The manual parsing records each speaker's name, chamber (including a flag for Extensions of Remarks), whether the speaker is referred to by title, as well as the starting and ending lines of each speech. To reduce the number of human errors, discrepancies between the manual and automated parsings are checked against the raw text, and errors from the manual parsing are corrected. A speaker's name is compared between manual and automated parsings after removing all whitespace and punctuation and coercing all characters to lower case.

The automated parser recognizes the beginning of a speech in the same location as the manual parsing in 94 (99.7) percent of speeches in the bound and daily editions, respectively. Of speeches with the same starting location not made by titled speakers, the parsings agree on the exact length of a speech in 88 (77) percent of cases and differ by two lines or fewer in 94 (92) percent of cases. Speeches with the same starting location also agree on the chamber in 99 (98) percent of cases and on the speaker's name in 98 (92) percent.

3.2 Matching

The audit of the speaker matching covers 138 (36) speeches from the bound (daily) edition of the *Congressional Record*. The speeches are chosen by a stratified sampling procedure that randomly selects two speeches from the automated parsing at the session-edition level. The two audits are treated sequentially: conditional on the set of automatically parsed speeches from the Record, this audit measures how well the algorithm matches speeches to members of Congress.

The automated and manual matchings agree in all but three cases for the bound edition and in all cases for the daily edition. The three disagreements arise when the manual mapping is made on contextual information in the *Congressional Record* that the parsing algorithm was not designed to recognize.

4 Variable Descriptions

Table 8 describes the names and values of variables for files in which they have at least one observation. Its keys are Variable and Files. Characters used as abbreviations and stand-ins are defined in Section 1.

Table 8: Variable Descriptions

Variable	Files	Type	Description
CHAMBER	manual/*/MMDDYYYY.txt	character	Chamber in which speech from manual parsing was made: H(ouse), S(enate), or E(xtensions of Remarks). See Table 7.
chamber	descr_*.txt, ###_SpeakerMap.txt	character	Chamber in which speech was made: H(ouse), S(enate), E(xtensions of Remarks), or N(one found). See Table 7 for descr and Section 2.4 for SpeakerMap.
char_count	descr_*.txt	numeric	Number of characters in a speech. See Section 2.2.
cong	hein-X.txt	numeric	Session of Congress.
context	hein-X_errors.txt	boolean	Indicator for whether a speech is manually matched with a congressperson using information beyond the standard demarcation. See Sections 2.2 and 3.2.
count	by*_2gram_###.txt	numeric	Count of bigram mentions. See Section 2.5.
date	descr_*.txt	numeric	Date on which speech was made. YYYYMMDD. See Table 7.
district	###_SpeakerMap.txt	numeric	Speaker's congressional district from historical source, if applicable. See Section 2.4.
end_line	audit_selections-hein-X.txt	numeric	Line of day-level file to stop manual parsing for audit. See Section 3.1.
file	descr_*.txt, errors_###.txt	character	Name of day-level file in which speech appears. See Section 2.1.
filename	audit_selections-hein-X.txt	character	Name of day-level file to manually parse for audit. See Section 3.1.
firstname	###_SpeakerMap.txt	character	Speaker's first name from historical source. See Section 2.4.
first_name	descr_*.txt	character	Speaker's first name from speaker demarcation. See Section 2.2.
gender	###_SpeakerMap.txt	character	Speaker's gender from historical source: M(ale), F(emale). See Section 2.4.
gender	descr_*.txt	character	Speaker's gender from speaker demarcation: M(ale), F(emale), Special (titled speaker), Unknown (not recorded). See Section 2.2.
lastname	###_SpeakerMap.txt	character	Speaker's last name from historical source. See Section 2.4.
last_name	descr_*.txt	character	Speaker's last name from speaker demarcation. See Section 2.2.

Table 8: Continued Variable Descriptions

Variable	Files	Type	Description
LINE_START	manual/*/MMDDYYYY.txt	numeric	Line of day-level file on which manually parsed speech begins. See Table 7.
line_start	descr_*.txt, errors_###.txt	numeric	Line of day-level file on which speech begins. See Table 7.
LINE_END	manual/*/MMDDYYYY.txt	numeric	Line of day-level file on which manually parsed speech ends. See Table 7.
line_end	descr_*.txt, errors_###.txt	numeric	Line of day-level file on which speech ends. See Table 7.
line_end_auto	errors_###.txt	numeric	Line of day-level file on which speech ends in automated parsing, if different from manual parsing. See Section 7.
line_end_manual	errors_###.txt	numeric	Line of day-level file on which speech ends in manual parsing, if different from automated parsing. See Section 7.
manual_speakerid	hein-X_errors.txt	numeric	Value of speakerid from manual matching of speeches to members of Congress. See Section 3.2.
mapped	hein-X.txt	numeric	Number of speeches matched with a congressperson in a session of Congress. See Section 2.4.
nonvoting	###_SpeakerMap.txt	character	Speaker's voting privileges from historical source. See Section 2.4.
number_within_file	descr_*.txt	numeric	Order of speeches in a day. See Table 7.
party	byparty_2gram_###.txt, ###_SpeakerMap.txt	character	Speaker's party from historical source. See Section 2.4.
phrase	by*_2gram_###.txt	character	Bigrams. See Section 2.5.
rate	hein-X.txt	numeric	Rate at which speeches are matched with a congressperson in a session of Congress. See Section 2.4.
session	audit_selections-hein-X.txt	numeric	Session to which day-level file belongs. See Section 3.1.
SPEAKER	manual/*/MMDDYYYY.txt	character	Manually parsed full speaker demarcation. See Section 2.2.
speaker	descr_*.txt, ###_SpeakerMap.txt	character	Full speaker demarcation. See Section 2.2.
speakerid	byspeaker_2gram_###.txt, ###_SpeakerMap.txt	numeric	Unique identifier for congresspeople within a chamber and session of Congress. Concatenation of the session of Congress, a unique 5-digit congressperson id, and the chamber of Congress (1 = Senate, 0 = House of Representatives).

Table 8: Continued Variable Descriptions

Variable	Files	Type	Description
SPECIAL	manual/*/MMDDYYYY.txt	character	Indicator (Y is yes, N is no) for whether the manually parsed speaker is referred to by title. See Section 2.2.
special	hein-X_errors.txt	boolean	Indicator (0 is no) for whether the speaker is referred to by title. See Section 2.2.
speech	speeches_*.txt	character	Full-text speeches following the cleaning step in 2.2
speech_id	descr_*.txt, hein-X_errors.txt, speeches_*.txt	numeric	Unique identifier for speeches within an edition of the *Congressional Record*. Concatenation of the session of Congress and order in which a speech appears.
start_line	audit_selections-hein-X.txt	numeric	Line of day-level file to start manual parsing for audit. See Section 3.1.
state	###_SpeakerMap.txt	character	Speaker's state from historical source. See Section 2.4.
state	descr_*.txt	character	Speaker's state from speaker demarcation. See Section 2.2.
type	procedural.txt	character	Bigram flag. See Section 2.6: co-occurring are phrases flagged by co-occurrence rules, riddicks are bigrams in *Riddick's Senate Procedure*, roberts are bigrams in *Robert's Rules of Order*, roberts_and_riddicks are bigrams appearing in both handbooks.
type	errors_###.txt	character	Reason for discrepancy between automated and manual parsing. See Section 3.1.
unmapped	hein-X.txt	numeric	Number of speeches not matched with a congressperson in a session of Congress. See Section 2.4.
word_count	descr_*.txt	numeric	Number of words in speech. See Section 2.2.
_classify	master_list.txt	character	Bigram flag. See Section 2.6 and the procedural type variable: bad_syntax is bad syntax, stopword is US-Congress-specific stopwords, and vocab is valid vocabulary.

Reference

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2017. Measuring polarization in high-dimensional data: Method and application to congressional speech. NBER working paper No. 22423.

Table 9: Manually Selected Stopwords

absent	adjourn	ask	can	chairman
committee	con	democrat	etc	gentleladies
gentlelady	gentleman	gentlemen	gentlewoman	gentlewomen
hereabout	hereafter	hereat	hereby	herein
hereinafter	hereinbefore	hereinto	hereof	hereon
hereto	heretofore	hereunder	hereunto	hereupon
herewith	month	mr	mrs	nai
nay	none	now	part	per
pro	republican	say	senator	shall
sir	speak	speaker	tell	thank
thereabout	thereafter	thereagainst	thereat	therebefore
therebeforn	thereby	therefor	therefore	therefrom
therein	thereinafter	thereof	thereon	thereto
theretofore	thereunder	thereunto	thereupon	therewith
therewithal	today	whereabouts	whereafter	whereas
whereat	whereby	wherefore	wherefrom	wherein
whereinto	whereof	whereon	whereto	whereunder
whereupon	wherever	wherewith	wherewithal	will
yea	yes	yield		

Appendix

A Processing the Vocabulary

The set of "valid bigrams"—those used for analysis in Gentzkow et al. (2017)—is a subset of those returned from the processing described in Section 2.2. Bigrams not spoken at least three times in at least one session are removed to ease the computational burden. To avoid double-counting, bigrams from the daily edition of the *Congressional Record* are ignored when bigrams from the bound edition are available. Bigrams are then selected for the final vocabulary by process of elimination.

First, bigrams with bad syntax are flagged. A bigram has bad syntax if it contains (i) any numbers, symbols, or punctuation; (ii) fewer than five characters, including the space; (iii) a one-letter word; or (iv) a word beginning with the first three letters of a month.

Second, bigrams containing the stem of a US-Congress-specific stopword are flagged. Stopwords come from three sources: (i) the manually selected stopwords in Table 9, (ii) the names of states, and (iii) the last names of all congresspeople recorded in the historical source.

Third, bigrams recording procedural speech are determined and flagged. These are bigrams that either directly appear in handbooks describing congressional procedure or frequently co-occur with the direct bigrams. The handbooks are *Robert's Rules of Order*, a widely accepted guide that explains the procedures of assemblies, and *Riddick's Senate Procedure*, a glossary-style document detailing the rules, practices, and customs of the United States Senate's operations and meetings.[6] Their text is parsed using a procedure similar to the one from Section 2.2. Each handbook is (i) coerced to lowercase, (ii) purged of hyphens and general English-language stopwords, (iii) broken into words by treating all non-alphanumeric characters as delimiters, (iv) stemmed, and (v) converted into a set of bigrams. Bigrams from *Robert's Rules of Order* (*Riddick's Senate Procedure*) are called Robert (Riddick) phrases. This notion is extended to speeches by defining a highly Robert (Riddick) speech as one for which Robert (Riddick) phrases account for at least 30 percent of all bigram counts. A similar notion classifies a speech as procedural when at least 30 percent of its bigrams appear in either handbook. The speech designations are not mutually exclusive.

Two co-occurrence rules are used to identify procedural phrases not in the handbooks. A phrase qualifies as procedural by the first rule if one of the following sets of conditions applies:

- It appears in at least 5 procedural speeches in more than 5 sessions and one of: 1) it appears in more than 5,200 highly Robert speeches, and at least 1.75 percent of speeches it appears in are highly Robert; or 2) it appears in more than 100 highly Robert speeches, and at least 7.5 percent of speeches it appears in are highly Robert; or 3) it appears in more than 50 highly Robert speeches, and more than 30 percent of speeches it appears in are highly Robert.

- It appears in at least 5 highly Robert speeches in more than 10 sessions and one of: 1) it appears in more than 2,000 highly Robert speeches, and at least 1 percent of speeches it appears in are highly Robert; or 2) it appears in more than 100 highly Robert speeches, and at least 5 percent of speeches it appears in are highly Robert; or 3) it appears in more than 50 highly Robert speeches, and at least 20 percent of speeches it appears in are highly Robert.

- It appears in at least 5 highly Riddick speeches in more than 10 sessions and one of: 1) it appears in at least 3,000 highly Riddick speeches, and at least 1.75 percent of speeches it appears in are highly Riddick; or 2) it appears in at least 100 highly Riddick speeches, and at least 7 percent of speeches it appears in are highly Riddick; or 3) it appears in at least 50 highly Riddick speeches, and at least 20 percent of speeches it appears in are highly Riddick.

Every phrase is also measured by the average percentage of Robert and Riddick phrases across speeches containing the phrase. Of the phrases not identified by the first rule, a phrase qualifies as procedural by the second rule if one of the following sets of conditions applies:

- 1) It is mentioned over 500 times; and 2) it appears in more than 5 sessions; and 3) over 5 percent of bigrams in speeches in which it occurs, on average, are Robert phrases.

[6] The 1876 version of *Robert's Rules of Order* was obtained from Project Gutenberg http://www.gutenberg.org/etext/9097 in early August 2009. A PDF of *Riddick's Senate Procedure* for the 101st session of Congress (1989–1991) was obtained from http://www.gpoaccess.gov/riddick/1441-1608.pdf on August 11, 2010 and converted to text using OCR with metadata removal.

- 1) It is mentioned over 20,000 times; and 2) it appears in more than 10 sessions; and 3) over 7.5 percent of bigrams in speeches in which it occurs, on average, are Riddick phrases.
- 1) It is mentioned over 500 times; and 2) it appears in more than 10 sessions; and 3) over 9.6 percent of bigrams in speeches in which it occurs, on average, are Riddick phrases.

The cut-off points for the two rules are chosen to maximize the share of excluded phrases, and minimize the share of non-excluded phrases, that are subjectively judged to be procedural.

Bigrams without a flag are treated as valid vocabulary. In cases when the three flags overlap, the procedural flag takes precedence over the flag for stopwords, which takes precedence over the flag for bad syntax.