**Gentzkow, Shapiro and Taddy: The Complete Congressional Record**

Gentzkow, Shapiro and Taddy (2019) have made available **all** speeches in the U.S. Congress from 1873 to 2017 (43rd to 114th Congress) as transcribed in the United States Congressional Record. They obtained the digital text from HeinOnline who performed Optical Character Recognition (OCR) on scanned print volumes. The digitized text of speeches as well as the meta information for the speeches (name of the speaker, together with his/her chamber/state/party affiliation, and the date of the speech) for each of the 72 consecutive sessions of Congress are available on their Stanford University website.

GST use the data to measure trends in the partisanship of congressional speech from 1873 to 2017. They define partisanship as the ease with which an observer can infer a congressperson’s party from his/her text utterances. Their findings are published in the July 2019 issue of Econometrica.

A tremendous amount of work was needed to make these speeches ready for digital use. The speeches span a very long uninterrupted period of 144 years and 72 sessions of Congress. Key meta variables for the speeches are available such as the date of the speech and the speaker's name, chamber affiliation (House or Senate), party affiliation, gender and home state. The organization of the content of the text/data files is described in the **codebook\_v2.pdf** on their Stanford website. The texts are fairly clean, considering the fact that the digitized speeches originate from the Optical Character Recognition (OCR) of scanned print volumes. OCR is challenging especially for the early sessions of Congress when the type-setting of old print is not particularly uniform and when the material is printed in more than one column. There are some errors in their files due to the fact that every now and then the scanned continual text is broken into incorrect speech segments; sometimes a unique speech is broken incorrectly into several speeches, and sometimes several distinct speeches are combined into a single speech. Also, there are occasional issues with mis-spelled and missing meta information. One must understand that it is virtually impossible to create such a large text/data set without errors, and the few problems mentioned here do not distract from the huge importance of this data base. GST are to be congratulated for their heroic effort.

The data on their website can be used to illustrate the tools that are discussed in this primer. An R program for reading the text information (the text of the speeches) and for merging it with the author (meta) information is listed on our book's website. We apply it to the speeches of the 43rd Congress (1873-1875). We illustrate three different ways of merging the text and the meta information. The **codebook\_v2.pdf** on the Stanford website describes the contents of the data files that are merged.

The file **speeches\_043.txt** contains the text of all speeches of the 43rd Congress (text, but without any meta information). Each row contains a speech identifier and the text of the speech. There are 119,302 speeches.

The file **043\_SpeakerMap.txt** contains the speaker information for a subset of these speeches. Each row contains the speech identifier, the speaker identifier (a number), the last and first name of the speaker (the listed last name has been cleaned of all misspellings and checked against the published list of House and Senate members), the chamber information (House/Senate), the home state, the gender (M/F), and the party affiliation; but note that this file does not contain the date of the speech. Not all speeches are listed; only those speeches (96,115) for which a clean speaker name could be attached.

The file speeches\_043.txt and the file 043\_SpeakerMap.txt are merged. The resulting file **speechesWithMeta1\_043.txt** contains the text and the speaker information for a subset of these speeches (96,115 speeches). Only those speeches that could be assigned an unambiguous speaker from either the House or the Senate are shown; speeches by the Speaker of the House and the (Senate) President are also omitted.

There is another way one could merge the text and the meta information. The file **descr\_043.txt** contains several descriptors of the speeches; the file has the same number of rows as the file speeches\_043.txt. It includes the speech identifier, the chamber origin of the speech (House/Senate), the date of the speech, the automatically extracted (and often misspelled) name of the speaker, a partially cleaned version of the last name of the speaker (with the label "unknown" if the extracted speaker could not be matched), and the speaker's home state (although quite often missing).

The file speeches\_043.txt and the file descr\_043.txt are merged. The resulting file **speechesWithMeta2\_043.txt** contains the text, the date of the speech, and the speaker information for **all** speeches (119,302) of the 43rd session. The partially cleaned last name (in variable speakercorr) is still not quite as clean as the corrected speaker names in the file 043\_SpeakerMap.txt and the merged file speechesWithMeta1\_043.txt. You can see this by looking at the frequencies of the corrected speaker names. Quite a few misspelled names are still remaining.

We prefer the first merged file **speechesWithMeta1\_043.txt** as it contains better speaker information. However, it does not contain the date of the speech. The date of the speech is included in the third merged file **speechesWithMeta3\_043.txt**.

**References**

Gentzkow, M, Shapiro, JM, Taddy, M: Measuring group differences in high-dimensional choices: Method and application to congressional speech. Econometrica, Vol 87 (July 2019), 1307-1340.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. <https://data.stanford.edu/congress_text>