**Performing Authenticity: Anti-Politically Correct Discourses in Brazil and the United States**

**Codebook**

Henrique Sposito

This codebook ranges from data collection to variable operationalization done for the article on *Anti-PC and Authenticity: Authenticity Performances in politics in Brazil and the United States*. All the data cleaning, preparation, analysis, and most of the data collection, was done through R software. For all of it, including the data, preparation, and analysis, refer to the “authenticity\_performances” GitHub repository. To access the repository, please request author’s permission/access. Besides that, the [poldis](https://github.com/henriquesposito/poldis) R package was developed to facilitate the wrangling and usage of political discourses beyond official speeches for analysis. The package is publicly available on GitHub[[1]](#footnote-1). The repository and package allow for a more transparent research process and easily reproducible findings. As follows, this codebook in divided in three sections. The first section covers data collection. The second section discusses data preparation. The third section clarify variables operationalization.

1. **Data Collection**

All the data on political discourses for the United States (US) was collected from the The American Presidency Project (APP) hosted by the University of California Santa Barbara., one of the largest and most complete databases for presidential documents. The data was scrapped directly in R using the rvest package (cite). The scripts for the scrapping are available, alongside the raw scraped data, in the data-raw folder of the “authenticity\_performances” repository. Six types of document categories were scraped from [The American Presidency Project](https://www.presidency.ucsb.edu/) (TAPP). These categories are debates, news conferences, interviews, speeches, miscellaneous oral remarks and campaign documents. These were scrapped alongside information on date, speakers and sources. Dates for the documents vary, but most categories were scrapped with 1960s as roughly the baseline year. It is clear, by looking at the data and for obvious technological reasons related, that data is far more abundant in the after the 1960s than before that. Table 1 (below) illustrates the number of observations scrapped for each document type earliest documents scraped and raw dataset name in the repository for the US. There is some overlap in documents scraped and raw data scope is broader than what the paper covers. For this reason, US discursive data scraped will be wrangled and cleaned before used (see next section). For the dataset scrapping scripts, please refer to the US\_scrapping script in the data-raw folder in the “authenticity\_performances” repository.

US Raw Data – Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Document Type** | **Total Observations** | **Earliest Date** | **Raw Dataset Name** | **Source** |
| Campaign | 3240 | 1976 | US\_Campaign | APP |
| Debates | 171 | 1960 | US\_Debates | APP |
| Interviews | 953 | 1864 | US\_Interviews | APP |
| Oral Remarks | 22010 | 1948 | US\_M\_Remarks | APP |
| Speeches | 5510 | 1972 | US\_Oral\_Remarks | APP |

When it comes to text data for Brazil, collection proved much more challenging in comparison to the US (see table 2 below). For oral remarks and official speeches, it relies on Cezar’s (2020) dataset which builds upon the archives for the Brazilian Presidential Library. The dataset includes all official oral remarks and speeches delivered by presidents since 1985. However, there are no large repositories or datasets for interviews or debates in Brazil openly available. Data for these were, thus, gathered by the author from subtitles for YouTube videos for these. There were three types of text collected from YouTube: debates, campaign remarks and interviews. On some occasions, pertinent videos that showed up in the searches that did not have subtitles in Youtube, but had many views, had their audio transcribed using an alternative software for which accuracy was also good[[2]](#footnote-2).

Regarding the usage of subtitles from Youtube videos, there are two considerations that must be discussed. One, these subtitles are automatically generated by software from the audio of video, this implies videos with worse audio quality may be subject to more misses (i.e. older videos). Two, on average, the automatic transcription generated is of good quality; however, it is expected that several nuances in language derived from uncommon words or related to a certain regional usages or accents might generate misses. Often, in both cases, software appears to omit the words instead of substituting them for others. Yet, in the absence of an alternative discursive database for Brazil, these present the best option to pursue to discourse analysis empirically considering different political settings in which politics “are done”.

For debates, the text for available runoff presidential debates since 1989, the year of the first direct presidential election in Brazil, were gathered for Brazil[[3]](#footnote-3). There are some data inconsistencies for some elections related to the fact that some cycles saw less debates as they were decided in the first round (e.g. 1998) while in other cycles circumstances did not allow certain candidates to participate in most debates (e.g. Bolsonaro 2018)[[4]](#footnote-4). For these instances, the participation of the top two candidates in the first-round of debates. All these possible issues are made explicit and are accounted for during the analysis process. Debates were manually coded into a dataset by the author. The observations were separated by the speakers and year. That is, for each debate there are two observations, one for each speaker, with all the candidate’s answers for that debate combined in the text variable. In total, 29 observations were gathered from 17 different debates. This is like how the US debates data will be prepared for analysis (next section), but since the BR\_debates dataset was manually coded by the author this process was done at the dataset assembly stage.

For campaign remarks, for each runoff candidate from starting in 1989, all the campaign speeches, videos and propaganda made on election years available in full and with subtitles from YouTube for a candidate are scraped and added into a dataset manually[[5]](#footnote-5). This is done by searching Youtube for the candidate’s name, respective years and campaign[[6]](#footnote-6). The availability of several compilations of campaign speeches, propaganda and campaign videos made by a large research project in Brazil proved important sources for gathering materials, especially important for earlier elections in Brazil[[7]](#footnote-7). For example, starting in 2010 political campaigns had their own YouTube channels with campaign content. That implies, as with debates, campaign remarks are more widely available and are of better quality for later election cycles. In some cases, text of campaign videos or speeches included other speakers or narration, these were when possible reduced to only get candidate’s spoken portions. For Bolsonaro who due to health issues was not able to campaign for many months in 2018, some of his “Facebook Lives” pre campaign were gathered as they often substituted campaign speeches. The election year is chosen as candidates are made official and allowed to publicly campaign officially in election years according to legislation in Brazil.

For interviews, for each runoff candidate since 1989, live interviews with all major or news channels in Brazil (Globo, Bandeirantes, Record, SBT, RedeTV and Cultura) occurring in the period between one year prior to elections and one year after election/tenure, and including tenure years for elected officials, were gathered into a dataset[[8]](#footnote-8). This is done by searching the name of the candidate, the word interview (“entrevista”) and the TV channel name on Youtube for videos and only selecting interviews for the years of interest for respective candidate[[9]](#footnote-9). These channels represent most interview sources. Still, recently, major written, oral or online communication channels in Brazil (UOL, Folha de Sao Paulo, Estadao, Veja and Jovem Pan) also produced video content with interviews for candidates and elected officials, these were also inserted in the dataset when found, though these were not explicitly searched for[[10]](#footnote-10). For some earlier dates before the 2000s, due to the scarcity of interview videos available on YouTube for Brazil, the timelines were expanded by to two years before election and two years after election or tenure. Often short portions of interviews were found, in these cases interview portions were only gathered if they were 5 minutes long or more uncut. As expected, interviews with elected candidates were often more available as for when these candidates were in office and around election times.

Brazil Raw Data – Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Document Type** | **Total Observations** | **Earliest Date** | **Raw Dataset Name** | **Source** |
| Campaign | 175 | 1989 | BR\_Campaign | Youtube |
| Debates | 29 | 1989 | BR\_Debates | Youtube |
| Interviews | 262 | 1987 | BR\_Interviews | Youtube |
| Oral Remarks | 6130 | 1985 | BR\_Oral | Cezar (2020) |

As a general note, the discrepancy in observations for the two cases follows data availability and time scopes. While for the US, data is widely available to be scrapped for longer periods of time, for Brazil, several datasets were often hand assembled by the author. This means, these datasets were already cleaned and wrangled at the time of assembly. More to the point, note that the observations and data referred in the US table above are the raw datasets included as raw data in the “authenticity\_performances” repository, these will be cleaned in the next section before their text are pre-processed for the analysis.

1. **Data Preparation**

After data collection, datasets needed to be cleaned and better organized for subsequent analysis. This is an important step for US data, since the period and range of discursive data scraped for the country differs a great deal. Data for Brazil was also manually coded into datasets and, consequentially, most of the cleaning has been performed at dataset assembly stage. As for example, debates for Brazil were already separated for speaker during collection. All these changes and steps taken to clean US data can be viewed in the data\_preparation\_US script in the data folder of the “authenticity\_performances” repository[[11]](#footnote-11).

Oral remarks and speeches data were merged into one dataset for the US. Duplicates were removed as there was some overlap between the datasets. Speakers outside the scope of the analysis, as first-ladies and secretaries, were removed. The earliest date for observations in the dataset was set at 1981, since this is when Ronald Reagan took office. The final dataset, US\_oral, was added to the data folder of the “authenticity\_performances” repository and contains 15016 texts (see also Table 3 below).

For interview data for the US, besides restricting the time for 1980 and onwards and selecting only speakers of interest, several interviews for elected officials and runner upper candidates were present in the US campaign data. This entailed that we selected all the text documents in the campaign data in which the word “interview” was present in title and merge with interviews dataset. The final dataset, US\_interviews, was added to the data folder of the “authenticity\_performances” repository and contains 936 texts (see also Table 3 below).

Campaign data for the US contained many speakers which we were not interested in as campaign managers and spokespersons. These were removed and only relevant speakers were maintained. As well, all observations with the word “interview” in titles were removed as they were added to interviews dataset. Moreover, all interviews with the words “press releases” in title were removed, as these were usually statements read by campaign staff. A few other observations with words as “commencement” and “campaign messages” in title were also removed as they appear to be misplaced (for commencement speeches) or simply happy holiday messages (and other short wishes) not delivered by speakers of interest. The final dataset, US\_campaign, was added to the data folder of the “authenticity\_performances” repository and contains 1563 texts (see also Table 3 below).

The debates dataset is where most of the working in cleaning the data happened. All primary debates (for party nominations) and vice-presidential debates were removed. Debates from 1980s and onwards were selected. Since this became a small dataset, speakers in each debate were manually reviewed, due to scrapping issues with data, and gathered. At this point, debates were split by speakers. This was done with the [get\_speaker()](https://github.com/henriquesposito/poldis/blob/master/R/get_speaker.R) function in poldis. That is, the text of each debate was separated by speaker in a separate data frame. Once separation was completed, speakers date and titles were recoded into the data frame. The final dataset, US\_campaign, was added to the data folder of the “authenticity\_performances” repository and contains 59 texts (see also Table 3 below). The 59 texts refer to speakers coded for 28 runoff debates in the US since 1980[[12]](#footnote-12).

US and Brazil Cleaned Data – Table 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Document Type** | **Observations** | **Earliest Date** | **Source** |
| U.S. | Speeches | 15016 | 1981 | TAPP |
| Campaign | 1563 | 1980 | TAPP |
| Debates [[13]](#footnote-13) | 59 | 1980 | TAPP |
| Interviews | 936 | 1980 | TAPP |
| Brazil | Speeches | 6130 | 1985 | Cezar (2020) |
| Campaign | 175 | 1989 | Youtube |
| Interviews | 262 | 1987 | Youtube |
| Debates | 29 | 1989 | Youtube |

At this stage, data was explored for insights. This included, for example, producing descriptive statistics as word frequencies, topic modelling, and sentiment analysis. These were done for datasets as a whole and/or for certain speakers for comparison examples. Though not very informative, this is helpful in investigating if setting really changes how speakers behave and to investigate variations in time. I will start with the US data, move to Brazil, and then compare both in some respects. All the information and work done at this section can be found in the data\_exploartion\_US and data\_exploration\_BR scripts in the data folder of the “authenticity\_performances” repository. These contain commented codes that should allow users to reproduce all the tables, figures and graphs included in this section.

1. **Variables’ Operationalization**

A definition is only as good as the dictionary (I would hope this is a famous saying). In this case, this means each authenticity performance is a s good as the “bag of words” defining it. For this reason, each performance needs to be well operationalized. Table 3 (below) describe how each authenticity performance was operationalized by the words and expression searched for. Please also refer to Table 1 in article for more info on performances and mechanisms. The approach taken here is to leave stop words in text as directionality is deemed important. That implies, the dictionary will have to be larger and that matches have a much higher probability of being real matches. For English and Portuguese, the expressions coded are usually 2 to 3 words to maintain directionality (pronoun/determinator and verb/adjective). The single words sometimes found in the Portuguese dictionary refer to verb conjugations that take a specific pronoun and, thus, pronoun need not be explicit. This is a choice that leans on the side of false negatives instead of false positives; that is, we prefer more accurate matches even if the number of matches are reduced. The codes below were generated by listening to randomly chosen speeches, debates and campaign remarks for the US and Brazil.

Note that plural or past options are often left out as these are captured in analysis if they entail an addition of letters at the end of the last word for expression. They are included in analysis even if dictionary does not explicitly mention them. Some accents in Portuguese were standardized to avoid ambiguity and increase matching possibilities (e.g. “é” became “e”, “nos” became “nós”). The codes are also meant to reduce the possibility of overlap, even as some authenticity performances can share a few similar displays, as for example truth telling or consistency. For politically incorrect terms in politics, the *Politically Correct Dictionary and Handbook* (Beard and Cerf 1992) was used as reference purposedly. The logic is to look for terms and expressions deemed politically incorrect in 1992 to establish a baseline for terms that are deemed incorrect for a long time and that politicians likely know about. Note also there is an effort to create dictionaries that are similar across languages in content and number of codes.

Dictionary of Terms for Authenticity Performances – Table 4

|  |  |  |
| --- | --- | --- |
| *Authenticity*  *Performance* | *Lexicon*  *English (Broad)* | *Lexicon*  *Portuguese (Broad)* |
| Truth Telling | “am telling the truth”, “are telling the truth”, “is telling the truth”, “the truth is”, “this is the truth”, “not lying”, “not lies”, “no lies”, “not telling you lies”, “is honest”, “am honest”, “is being honest”, “are being honest”, “are honest”, “honesty”, “is sincere”, “are sincere”, “am sincere”, “is being sincere”, “are being sincere”, “is true”, “are true”, “not a liar”, “bottom of my heart”, “I swear”, “I reassure”, “we reassure”, “I assure”, “we assure”, “be assured”, “is truthful”, “are truthful”, “am truthful”, “is being truthful”, “are being truthful”, “I know that”, “is evident”, “are evident”, “I am sure”, “trust me”, “am frank”, “are frank”, “is frank”, “being frank”, “is upfront”, “are upfront”, “am upfront”, “being upfront”, “will come clean”, “am coming clean”, “are coming clean”, “is straightforward”, “are straightforward”, “being straightforward”, “believe me”, “I am certain”, “no bullshit”, “not bullshitting”, | “a verdade e”, “esta e a verdade”, “digo a verdade”, “dizemos a verdade”, “pura verdade”, “não e mentira”, “não estou mentindo”, “e honesto”, “sou honesto”, “somos honesto”, “sendo honesto”, “a honestidade”, “ser sincero”, “e sincero”, “com sinceridade”, “e verdade”, “são verdadeiras”, “não sou mentiroso”, “não minto”, “fundo do meu coração”, “sou verdadeiro”, “somos verdadeiros”, “tenho certeza”, “certeza absoluta”, “confia em mim”, “confie em mim”, “pode confiar”, “sou franco”, “somos francos”, “fraqueza”, “falando a verdade”, “falo a verdade”, “falamos a verdade”, “acredite em mim”, “pode acreditar”, “podem acreditar”, “eu tenho certeza”, “isso e a verdade”, “somos honestos”, “com honestidade”, “toda a sinceridade”, “com sinceridade”, “toda sinceridade”, “sou confiável”, “somos confiáveis”, “as coisas são assim”, “a realidade das coisas”, “juro por deus”, “com certeza”, “digo com precisão”, “veracidade”, “premissa”, “afirmo para vocês”, “isso e como aconteceu”, “falar umas verdades” |
| Lying Accusing | “not truth”, “not the truth”, “not true”, “aren’t true”, “isn’t true”, “being untruthful”, “is lying”, “are lying”, “is a liar”, “are liars”, “is dishonest”, “are dishonest”, “being dishonest”, “is fake”, “are fake”, “being fake”, “is corrupt”, “are corrupt”, “full of lies”, “not sincere”, “not being sincere”, “isn’t sincere”, “aren’t sincere”, “not honest”, “not being honest”, “is cheating”, “is a cheater”, “are cheaters”, “are cheating”, “are tricking”, “is tricking”, “be deceived”, “is deceiving”, “are deceiving”, “are a hypocrite”, “is a hypocrite”, “are being a hypocrite”, “is being a hypocrite”, “is crooked”, “are crooked”, “is misleading”, “are misleading”, “has double-standards”, “are sneaky”, “is sneaky”, “ has two faces”, “two-faced”, “has double faces”, “double-faced”, “you are wrong”, “not correct”, “fooled by”, “do not believe”, “is misrepresenting”, “they misrepresent”, “is misrepresent”, “are misrepresent”, “pretends that”, “pretends to”, “is pretending”, “are pretending”, “keep pretending”, “breach your trust”, “breach of trust”, “is false”, “are false”, “being false”, “is misinforming”, “are misinforming”, “being misinformed”, “pretended”, “cut the crap”, “full of crap” | “não e verdade”, “não e verdadeiro”, “e mentiroso”, “está mentindo”, “são mentiroso”, “e mentira”, “de mentira”, “tudo mentira”, “e desonesto”, “mentiram”, “mentiu”, “um desonesto”, “esse desonesto”, “de desonesto”, “são desonesto”, “e falso”, “são falsos”, “são corruptos”, “e corrupto”, “de corrupto”, “todos corrupto”, “não são sincero”, “não e sincero”, “não são honestos”, “não e honesto”, “são trapaceiros”, “e trapaceiro”, “eles trapaceiam”, “trapaceou”, “e enganar”, “ser enganado”, “vão enganar”, “sendo enganados”, “e hipócrita”, “e enganador”, “e enganação”, “duas caras”, “enganado por”, “não acredite”, “eles finge”, “ele finge”, “e fingimento”, “ela finge”, “quebrou a sua confiança “, “quebra de confiança”, “ e falso”, “são falsos”, “falsidade”, “e ficção”, “história para boi dormir”, “historinha para boi dormir”, “e calunia”, “são calunias”, “difamação”, “difamar”, “uma inverdade”, “são inverdades”, “e inverdade”, “isso e invenção”, “essas são invenções”, “isso e uma lenda”, “essas são ledas”, “tenta iludir”, “tentando iludir”, “uma farsa”, “tramoia”, “mal intencionado”, “mas intenções”, “falta de informação”, “esta mal-informado”, “estão mal-informados” |
| Consistency | “we delivered”, “I delivered”, “check and see”, “I keep my word”, “we keep our word”, “I kept my word”, “we kept our word”, “I keep my promise”, “I kept my promise”, “we keep our promise”, “as promised”, “we kept our promise”, “am responsible”, “I take responsibility”, “we take responsibility”, “we assume responsibility”, “we are accountable”, “we are responsible”, “our duty”, “my duty”, “give my word”, “giving my word”, “own up my”, “owning up my”, “accept responsibility”, “accept the blame”, “recognize my mistakes”, “admit I was wrong”, “I made mistakes”, “I guarantee”, “we guarantee”, “I can guarantee”, “we can guarantee”, “I promise”, “we promise”, “we can prove”, “I can prove”, “we proved”, “I proved”, “am reliable”, “rely on me”, “rely on us”, “be reassured”, “you can hold me accountable”, “you can hold us accountable”, “see with your own eyes”, “vote of confidence”, “our mission”, “my mission”, “my commitment”, “our commitment”, “during our government”, during my government”, “while I was in charge” | “nós entregamos”, “eu entreguei”, “veja com seus próprios olhos”, “cumpro minhas palavras”, “cumprimos nossas palavra”, “cumpri minha palavra”, “cumpro minhas promessas”, “nossas promessa”, “um compromisso”, “meu compromisso”, “tenho um compromisso com”, “eu sou responsável”, “eu assumo a responsabilidade”, “nós somos responsáveis”, “nós assumimos a responsabilidade”, “nosso dever”, “meu dever”, “dou minha palavra”, “faço uma promessa”, “fazer uma promessa”, “aceitar a responsabilidade”, “aceito a responsabilidade”, “aceitamos a responsabilidade”, “aceitar a culpa”, “meus erros”, “que errei”, “eu errei”, “eu garanto”, “eu posso garantir”, “eu prometo”, “podemos provar”, “posso provar”, “provaremos”, “eu provei”, “voto de confiança”, “encarrego pessoalmente”, “encarreguei pessoalmente”, “estou comprometido”, “meu comprometimento”, “comprometimento com”, “o comprometimento”, “fazer o possível”, “minha supervisão”, “minha missão”, “nossa missão”, “no meu governo”, “no nosso governo”, “durante nosso governo”, “eu era encarregado”, “eu era o encarregado”, “fomos encarregados de” |
| Finger pointing | “are inconsistent”, “is inconsistent”, “being inconsistent”, “are irresponsible”, “is irresponsible”, “being irresponsible”, “their fault”, “not my fault”, “not our fault”, “they left us with”, “they are responsible”, “are not responsible”, “aren’t responsible”, “is not responsible”, “isn’t responsible”, “costed us”, “false promises”, “lack accountability”, “lacking accountability”, “not kept their word”, “not kept his word”, “not kept her word”, “not kept promises”, “not kept the”, “not kept his”, “not kept her”, “not kept their”, “not keep their word”, “not keep his word”, “not keep her word”, “not keep the”, “didn’t keep the”, “didn’t keep her”, “didn’t keep his”, “hasn’t kept his”, “hasn’t kept her”, “not recognize”, “he made mistakes”, “she made mistakes”, “they made mistakes”, “not our mistake”, “not my mistake”, “not take responsibility”, “not my responsibility”, “not accountable”, “him accountable”, “them accountable”, “her accountable”, “blame them”, “blame him”, “blame his”, “blame her”, “their blame”, “break promises”, “broken promises”, “has betrayed”, “they betrayed”, “betraying”, “will betray”, “has tricked”, “has lied”, “not deliver”, “didn’t deliver”, “hasn’t deliver”, “failed your obligations”, “failed in your obligations”, “failed his obligations”, “failed her obligations”, “failed in his duty”, “failed in her duty”, “failed his duty”, “failed her duty”, “failed your duties”, “stabbed in the back” | “e inconsistente”, “são inconsistente”, “e irresponsável”, “são irresponsáveis”, “culpa deles”, “a culpa não e minha”, “não e minha culpa”, “eles nós deixaram”, “são responsáveis”, “e responsável”, “nós custou”, “falsas promessas”, “falta de prestação de contas”, “falharam”, “falhou”, “não cumpriu”, “não cumpriram”, “não reconheceu”, “não reconheceram”, “errou”, “erraram”, “não se responsabiliza”, “não me responsabilizo”, “culpa e sua”, “sua culpa”, “quebrar promessas”, “promessas quebradas”, “quebra de promessas”, “fala uma coisa e faz outra”, “fala uma coisa aqui e faz outra”, “falsas promessas”, “são trapaceiros”, “cometeu erros”, “cometeram erros”, “não reconhece”, “não reconheceu”, “assumiu a responsabilidade”, “promete uma coisa”, “promete o mundo”, “traiu a confiança”, “traiu a sua confiança”, “quebra de confiança”, “quebraram sua confiança”, “e falcatrua”, “foi falcatrua”, “cheio de falcatrua”, “houve fraude”, “houveram fraudes”, “fraudulento”, “uma negociata”, “facada nas costas”, “faltou com respeito”, “não faz o que promete”, “não fez o que promete”, “promessas em vão”, “palavras em vão”, “falta de comprometimento”, “falta de compromisso”, “houveram desvio”, “houve desvio”, “a culpa e do”, “cheio de promessas”, “a conta não fecha”, “não terminaram” |
| Origins | “I was born”, “I come from”, “we come from”, “I grew up”, “growing up in”, “my parents”, “my mom”, “my mother”, “my father”, “my dad”, “my family”, “raised me”, “I was raised”, “we were raised”, “we grew up”, “my background”, “being surrounded by”, “being exposed to”, “my siblings”, “going to school in”, “our local church”, “Sunday mass”, “Saturday mass”, “family tradition”, “tradition in my house”, “in our house”, “growing up”, “back in the day”, “my grandparents”, “in my town”, “in my state”, “in my region”, “our community”, “in my community”, “our town”, “our state”, “my hometown”, “our hometown”, “my home state”, “our home state”, “back home”, “our house”, “my house”, “our neighbourhood”, “in my district”, “I lived in”, “we lived in”, “we used to play”, “I used to play”, “I was thought” | “Eu nasci”, “Eu vim de”, “eu venho de”, “viemos de”, “cresci”, “nós crescemos”, “meus pais”, “minha mãe”, “minha mãe”, “minha família”, “fui criado”, “fomos criados”, “minhas origens”, “meus irmãos”, “meu irmão”, “minha irmã”, “tradição familiar”, “tradição em casa”, “crescendo”, “antigamente”, “meu avô”, “minha avó”, “meus avós”, “na minha cidade”, “no meu estado”, “na minha região”, “nossa comunidade”, “na minha comunidade”, “nossa cidade”, “nosso estado”, “cidade natal”, “estado de origem”, “minha casa”, “nossa casa”, “lá em casa”, “nosso bairro”, “no meu bairro”, “eu morava”, “vivíamos”, “na minha terra”, “de onde eu venho”, “missa de domingo”, “missa toda semana”, “brincava”, “eram outros tempos”, “fui educado”, “morávamos”, “eu morei”, “nós moramos”, “de onde venho”, “eram tempos diferentes” |
| Common Sense | “is common sense”, “are common sense”, “everyone knows”, “it is undeniable”, “stating the obvious”, “say the obvious”, “everyone agrees”, “we all know”, “common wisdom”, “the people know”, “popular knowledge”, “from experience”, “it is my experience”, “sound judgment”, “practical solution”, “practical choice”, “practical answer”, “pragmatic solution”, “pragmatic answer”, “pragmatic choice”, “realistic answer”, “let me tell you about”, “is obvious”, “are obvious”, “obvious answer”, “obvious solution”, “as we all learned”, “we have all learned that”, “do not need to tell you that”, “the reality is”, “there is no logic”, “it does not make sense”, “it doesn’t make sense”, “we know it does not work”, “no one disagrees that”, “no person disagrees”, “there is not a person”, “there is not a human being”, “there is not a family”, “there is not an American”, “there is no single citizen”, “there is not one single person”, “there is not one single human being”, “there is not one single family”, “there is not one single American”, “there is not one single citizen”, “there is not one single person”, “there is not one human being”, “there is not one family”, “there is not one American” | “senso comum”, “bom senso”, “todos sabem”, “afirmando o óbvio”, “todos concordam”, “todos sabemos”, “sabemos todos”, “todos nós sabemos”, “sabedoria popular”, “por experiência”, “e minha experiência”, “sou prático”, “tem que ser prático”, “devemos ser prático”, “sendo prático”, “sou pragmático”, “tem que ser pragmático”, “devemos ser pragmático”, “sendo pragmático”, “sou realista”, “sendo realista”, “sejamos realista”, “realisticamente falando”, “e óbvio”, “como todos nós aprendemos”, “como sabemos”, “não preciso te dizer”, “o povo sabe”, “agente aprendeu”, “nós aprendemos”, “nós sabemos”, “não tem logica”, “como aprendemos”, “não faz sentido”, “não fazem sentido”, “estamos cansados de saber”, “sabemos que não funciona”, “ninguém discorda que”, “não tem uma pessoa”, “não existe uma pessoa”, “não há uma pessoa”, “não existe um ser humano”, “não tem um ser humano”, “não há um ser humano”, “não tem uma família”, “não existe uma família”, “não há uma família”, “não tem um brasileiro”, “não há um Brasileiro”, “não existe um brasileiro”, “não tem uma brasileira”, “não há uma Brasileira”, “não existe uma brasileira” |
| Anti-PC | “politically correct”, “political correctness”, “PC”, “plain speaking”, “speaking my mind”, “speak my mind”, “say what I think”, “saying what I think”, “not going to pretend”, “not pretend”, “speak what you think”, “not what you want to hear”, “not butter up”, “not beat around the bush”, “cut to the chase”, “just being real”, “saying what everyone thinks”, “say what everyone is thinking”, “speaking plainly”, “coloured people”, “negro”, “retarded”, “nigger”, “third world”, “oriental people”, “crippled people”, “is crippled”, “culturally deprived”, “drug addict”, “junkie”, “drunk”, “fat people”, “fat person”, fat population”, “handicapped”, “homosexual” “faggot”, “deviant”, “perverted”, “illegals”, “illegal immigrants”, “illegal alien”, “Jew”, “non-white”, “prostitutes”, “promiscuous”, “stupid”, “tribe”, “underdeveloped” | “politicamente correto”, “falar francamente”, “falando francamente”, “falar o que penso”, “falo o que penso”, “falando o que penso”, “dizer o que penso”, “papas na língua”, “não vou fingir”, “não estou aqui para agradar”, “falar o que você pensa”, “o que você quer ouvir”, “não adulterar”, “não rodeio”, “não dou rodeio”, “direto ao ponto”, “dizer o que todos pensam”, “dizendo o que penso”, “dizendo o que todos pensam”, “dizer o que todos estão pensando”, “não vou amaciar”, “não dá para amaciar”, “gordos”, “retardado”, “retardada”, “veado”, “população preta”, “os pretos”, “as pretas”, “terceiro mundo”, “viciado em drogas”, “bêbado”, “drogado”, “sem cultura”, “pervertidos”, “promíscuo”, “imbecil”, “estupido”, “aleijado”, “defeituoso”, “incapacitado”, “inválido”, “mongoloide”, “deficiente mental”, “deficiência mental”, “o incapacitado”, “a incapacitada”, “travesti”, “homossexualismo” |
| Territorial | “have been to”, “have visited”, “came all the way to”, “back from”, “will visit”, “saw first-hand”, “see first-hand”, “we visited”, “I visited”, “we visited”, “travelled to”, “traveling to”, “spend a few days in”, “spent some time in”, “spent time in”, “met great people in”, “we were hosted”, “I was hosted”, “our time in”, “my time in”, “our visit”, “spent a lot of time in”, “were many times in”, “got to know the whole country”, “got to know all the states” | “estive em”, “visitei”, “voltou de”, “voltei de”, “voltando de”, “voltamos de”, “estive em”, “estivemos em”, “visitará”, “visitarei”, “vi em primeira mão”, “ver em primeira mão”, “visitamos”, “viajei para”, “passei alguns dias em”, “passei algum tempo em”, “passei um tempo”, “conheci ótimas pessoas”, “conhecemos ótimas pessoas em”, “fomos hospedados”, “minha passagem”, “nossa passagem”, “nossa visita”, “fui muitas vezes para”, “estive muitas vezes em”, “passei muito tempo em”, “meu tempo em”, “estive por todo o Brasil”, “de norte a sul do pais”, “conheço todo o pais”, “conheci todo o pais”, “conheci todo o Brasil”, “conheço todo o Brasil” |

1. To download the package from R in your system, please make sure the package devtools is installed and then run the following line: devtools::install\_github(“henriquesposito/poldis”). [↑](#footnote-ref-1)
2. The software used in these occasions is called [happyscribe](https://www.happyscribe.com/). [↑](#footnote-ref-2)
3. The runoff elections (second round) were chosen since in the first round there are often too many candidates debating which only allow speakers to talk for a few minutes in total. While, at the same time, candidates in the runoff debates can speak for much longer times. Tentatively, at least two debates for each election since 1989 were gathered. For the 1994 election, the dataset contains only one debate because the election was decided in the first round, with Cardoso getting over 50% of the votes. The debate happened in the first round of elections and was the only national debate where all major candidates present were present for the election cycle. The two most voted candidates were coded in that case, Cardoso and Lula. For the 1998 election, also decided in the first round with the re-election of Cardoso, there were no debates as candidates were too many and did not agree on debating. For 2002, Lula won the election in the second round, however, the only debate found in full and with subtitles available is for the first round. In this case, the debate was coded so the two first candidates that went to second round of elections, Serra and Lula. For the second runoff debate of 2006 subtitles were not available, but complete video was. In that case, other software was used to automatically transcribe audio. [↑](#footnote-ref-3)
4. In 2018 Bolsonaro did not participate in most debates due to health issues related to the attack suffered early on in his campaign. At the same time, the runner upper, Fernando Haddad only participated in a few of the later debates. As a result, there were no debates in the second round and the two candidates never debated each other in the first round. Bolsonaro’s participation in the two first debates of the first round in 2018 are gathered for the dataset. For Haddad, his participation in the two last debates for the first round were gathered.. [↑](#footnote-ref-4)
5. For the 1994 and 1998 election cycles, since there was no runoff election, the campaign rallies of Lula were gathered as he was by far the second most voted candidates. [↑](#footnote-ref-5)
6. Four searches were done for each candidate looking one for name and election year, one with name, election year and the word campaign (“campanha”), one for name and election year, one with name, election year and the word speeches (“discurso”), and one for name and election year, one with name, election year and the word rally (“comicio”). For some early years, campaign speeches are hardly available, but campaign videos are. [↑](#footnote-ref-6)
7. See for example the research project at the State University of Rio de Janeiro (UERJ) for certain election cycles and made available publicly on their YouTube [channel](https://www.youtube.com/channel/UCkcuDdIEuQ9YqOjHsp4-EHQ/videos). This means many observations for certain year and speaker are grouped together, this is not an issue for the type of analysis proposed since these discourses will be subset at the unit of analysis, authenticity performances. [↑](#footnote-ref-7)
8. For these, only the interviews in full and where subtitles are available were scrapped. It is a tradition that presidential candidates and presidents in Brazil attend an interview show with several journalists from different media channels called “Roda Viva” at least once in Cultura, therefore, the channel though not major was added to the sample. For both Serra and Cardoso, interviews with Roda Viva available were much outside the time scope looked for and were left out for this reason. For Globo and Bandeirantes, both channels have news channels (Globo News and Band News) which were also considered as part of the major channel and videos found for these were added to the dataset. [↑](#footnote-ref-8)
9. For Collor, because interviews are scarce for that time period on YouTube and due to his impeachment process, observations for a longer time period from 1987 to 1995 were gathered. For Lula, his 1986 Roda Viva interview was also added as it is one of the few interviews for the pre-1989 campaign period. [↑](#footnote-ref-9)
10. Very recently CNN has funded a more accessible Brazil branch, videos from this source were also added if they were found. As with other online and written channels, CNN was not searched for. Moreover, Bolsonaro is known for giving interviews to smaller online blogs and news channels. In some cases, if these appeared in the searches and had over 1 million views they were added to the dataset. [↑](#footnote-ref-10)
11. The script details all the steps taken to prepare the data. Users can also run the script to reproduce steps. [↑](#footnote-ref-11)
12. Usually, runoff debates in the US only have 2 candidates (with rare exceptions where independent candidate is invited) so for each debate at least 2 speakers were coded. For Brazil, the two runoff candidates in the 2018 election did not debate each other due to Bolsonaro’s health condition, in that case the participation of Bolsonaro in the 2 first debates of 2018 and of Haddad in the following 3 in the first round were gathered. [↑](#footnote-ref-12)
13. The number of debates for both Brazil and the US reflect the number of speakers coded for each runoff debate in dataset. That is, the text of each debate was separated by speaker. In the US, runoff debates usually have 2 candidates (with rare exceptions where independent candidate is invited). For Brazil, the two runoff candidates in the 2018 election did not debate each other due to Bolsonaro’s health condition, in that case the participation of Bolsonaro in the 2 first debates of 2018 and of Haddad in the following 3 in the first round were gathered. [↑](#footnote-ref-13)