CA

Journal of **Cultural Analytics**

TED Talks as Data

Katherine M. Kinnaird and John Laudun

07.19.19

Peer-Reviewed By: Ian Milligan

Clusters: Data

Article DOI: 10.22148/16.042

Dataverse DOI: https://github.com/kinnaird-laudun/data/tree/master/Release_v0

Journal ISSN: 2371-4549

Cite: Katherine M. Kinnaird and John Laudun, "TED Talks as Data," Journal of Cultural Analytics. July 19, 2019.

> A group of teenagers cluster near their lockers, enjoying quick conversations between classes. One of them goes a little too long and, realizing it, addresses the group and the situation by announcing, "Well, thanks for coming to my TED talk." The rest laugh, nod their heads, and the conversational flow returns to normal before the bell sounds announcing that classes are about to begin. (From field notes by one of the authors.)

Over the past decade, TED talks have achieved a high level of cultural currency. In addition to the original event, which continues, the various spinoff events, e.g. TEDMed or TEDGlobal, and the localized versions (TEDx), there is a TED radio hour on National Public Radio as well as a considerable infrastructure designed to formalize the inclusion of TED talks in education. The nature of a TED talk has so saturated American culture that one of us has documented, in fact, the saying noted above offered when someone has monopolized conversation too long: "Thanks for coming to my TED talk."[1]

[1] Part of this research was performed while the authors were visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation, as part of

Such an ironic notion of a TED talk relies on audience familiarity with the original, and, like other uses of irony and parody, reveals an intuitive grasp of the nature of a thing or form. One indication of the maturity of the genre is that even parodies of the form of most TED talks have themselves been given as talks.[2] The emergence of a stable form, TED talks as a genre, paired with the perceived success of that form has even elicited guides and coaches.[3] our own interest is in understanding the nature of TED talks not only in terms of what makes a talk popular, but how do topic, gender, discipline, and vocabulary manifest themselves in a TED talk, and are TED talks as distinct as parodies suggest? That is, can we build a computational model of a TED talk? Additionally, can such models help discern patterns of influence among speakers and topics?

In the process of doing such work, we found ourselves downloading, parsing, collating, and compiling a fair amount of data that in the interest of advancing culture analytics theories and methods, this paper introduces a collection of cleaned and collated data sets of TED talk transcripts, including meta-information about each talk and relevant speakers. We contextualize the utility of having such data sets by considering TED talks as a particular kind of cultural artifact: one that offers us a chance to open up the inventory of human textual production, that confronts us with the status of such artifacts, and that can be added to the archaeological record. The remainder of the paper is organized as follows. In the next section, we offer three perspectives on TED talks and their transcripts as data, as cultural artifacts, and as examples of human oral tradition. After our consideration of the dimensions of the data set, in the following section we introduce how the TED talk data set was scraped and cleaned from the raw HTML files and then preview a selection of the data set. In the final section, we summarize our contributions and preview avenues for future work using this newly compiled data set.

the Cultural Analytics Long Program. The first author is the Clare Boothe Luce Assistant Professor of Computer Science and Statistical and Data Science at Smith College and as such, is supported by Henry Luce Foundation's Clare Boothe Luce Program. The second author was supported by Dr. Doris Meriwether/BoRSF professorship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

[2] Will Stephen. *How to sound smart in your TEDx Talk.* https://youtu.be/8S0FDjFBj8o. 2014.

[3] Peter Smith. *How to Give a TED Talk.* https://www.youtube.com/watch?v=ohZk7pPJN8o. 2012; Fia Fasbinder. *Want to Get on the TED Stage? Here's How.* https://www.inc.com/fia-fasbinder/how-do-you-get-on-the- ted-stage-do-this-first.html. 2017.) Building off previous work on TED talks that has mainly focused on delineating what makes a "good" talk, usually measured in terms of popularity,((June Cohen. *What Makes A Great TED Talk.* https://youtu.be/RVDfWfUSBIM. 2010; Sebastian Wernicke. *Lies, damned lies and statistics (about TEDTalks).* http : / / www . ted . com / talks / lies _ damned _ lies _ and _ statistics_about_tedtalks. 2010.

2. The Nature of the Data Set

Conceived in the same year as, and featuring demonstrations of, the compact disc and the Apple MacIntosh, the first TED conference in 1984 included presentations by Benoit Mandelbrot, Nicholas Negroponte, and Stewart Brand. It was six years before the next instance of the conference, but during that time, according to the TED mythos, the world came around. During the nineties, TED expanded beyond its original trinity of "technology, entertainment, and design" and, in the process, built a broader base of supporters, many of whom became regular attendees, sometimes known as "TEDsters." As the decade closed, TED's original curator reached out to Chris Anderson, who had himself begun a similarly diverse media enterprise, Future, in 1985. By the end of 2001, Anderson had taken over the running of the conference. By 2005, recognizing the expanding catalog of content that TED had accumulated over the years, the conference that was now increasingly a multimedia organization began offering a selection of those talks which had received the highest ratings in video form. From the start, the videos were offered under the Creative Commons Attribution-NonCommercial-NoDerivs (BY-NC-ND). Within six months, the 44 talks published had been viewed more than three million times. With that success, the TED organization invested in expanding the online offerings. Its efforts were rewarded with a number of awards, culminating with a Peabody Award in 2012.[4]

The other, named TED conferences have diverse origins and developments. The first of these was TEDMED, started in 1998 by TED's founder Richard Wurman and was restarted in 2008 under new ownership. In 2010, TEDMED talks became a part of the offerings on TED website. Anderson started TED Global in 2005 with the idea of having an internationally focused series of conferences held in a different location around the world each year. TEDWomen began in 2010 as a three-day conference and continues to the present. TEDYouth began the following year in 2011 with a focus on middle and high school students and a mix of speakers and hands-on activities. In addition to these series, there is of course also TED-Ed and the TED Radio Hour, both launched in the spring of 2012. Finally, there is TEDx, those talks which operate under the umbrella of TED by following the format for speakers and using its branding. In return, organizers

[4] TED itself maintains a robust collection of "about" pages. In compiling this brief history of TED and its products, we have drawn upon not only the organization's own account (https://www.ted.com/about/our- organization/history-of-ted) as well as Anderson's own TED talks, like the one he gave in 2002 (Chris Ander- son, "TED's Nonprofit Transition," TED 2002 [https://www.ted.com/talks/chris_anderson_shares_his_vision_ for_ted]) but also a number of write-ups about TED as it gained greater visibility and which, in turn, gave it more visibility. See for example this Forbes' report from 2012: "Here's Why TED and TEDx are So Incredibly Appealing"(Mark Fidelman. *Here's Why TED and TEDx are So Incredibly Appealing*. 2012.

of TEDx events agree not to pay speakers, only to collect fees to offset conference costs, and to relinquish copyrights to materials to TED.

Given the span of time, the stops and starts, the discovery of new audiences and new platforms, the original 18-minute format of TED talks has been flexed in many ways. While our sense might be, based on anecdotal information and experience—which is often reinforced by parodists and consultants, that TED talks are stable, we see this as something to be explored. The dataset presented in this paper offers one avenue for such exploration. If changes in scope or form are detectable, one of the questions then becomes how such changes are discerned, or not discerned, by TED's various audiences.

TED talks can be viewed through several lenses, as cultural artifacts, as data, as performances, as examples of human speech. In "TED Talks as Texts" (2.1), we explore the interesting opportunity that TED Talks present as objects of cultural study. In "TED Talks as Talk" (2.2), we discuss the nature of TED talks as highly composed but orally presented texts and how their hybrid nature presents some unique difficulties as well as some compelling opportunities for a consideration of textuality. In "TED Talks as Data" (2.3), we preview elements of our data set, which includes not only the talks but who gave the talk, when and where it was given, how many times a talk has been watched, as well as how the talk was tagged and described by the TED organization.

2.1 TED Talks as Texts

Current efforts to model human textual productivity using computational methods have mostly focused, to date, on either very large texts or very small ones. In the digital humanities, a range of computational studies focused mostly on novels have begun to explore the discernibility of genres and trends.[5] In the information and social sciences, small texts (like tweets) and specific social interactions (like re-tweeting and following) have mostly stood in as objects to trace other kinds of phenomena, be they cascades through social networks or the networks themselves.[6] This divergence in focus has meant that considerations of texts has

[5] Matthew Jockers. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013; Andrew Piper. *Enumerations: Data and Literary Study*. University of Chicago Press, 2018; Christof Schoch. *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama*. 2017. doi: {http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html}; Ted Underwood. "The Life Cycles of Genres". In: *Cultural Analytics* 23 (2016). http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

[6] PS Dodds, KD Harris, IM Kloumann, CA Bliss, CM Danforth (2011), "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PLoS ONE* 6 (12):

tracked from the rather large productions of a relatively small number of individuals to the rather small productions of the rest of humanity in the form, largely, of social media posts. There is also, we should note, differences in accessibility and in ontology of texts: the study of the large texts is circumscribed at one end of their history by efforts to digitize physical objects and, at the other end, by the entanglements of copyright; the study of small texts is mostly of "born digital" objects that are both numerous and easily available.

While there have been a number of interesting studies of what might be termed "texts in the middle" like reviews of books and movies drawn from sites either dedicated to such endeavors or from more general purpose fora, e.g. Reddit, we feel there is more work to be done on and among middle-sized texts.[7] These mesoscale texts, ranging from a few dozen words to a few thousand words, have been the principle products of verbal behavior for most of human history.[8] Many of these texts can be, and have been, classified under familiar names: myths, legends, tales, jokes, anecdotes, to name but a few genres long the purview of folklorists and anthropologists interested in how humans shape their views of the world, and thus the world itself, through stringing words together.

Accessible corpora of such texts are, however, few and far between. As much as folklorists have relied upon indices of tale types and motifs (smaller units of discourse), neither the indices nor the texts upon which they are built have transitioned into digital forms that can be easily queried or called upon for computational analysis.[9] Wanting to have texts that had some dimension of orality to

e26752. https://doi.org/10.1371/journal.pone.0026752; Filippo Menczer, "The Spread of Misinformation in Social Media," in *Proceedings of the 25th International Conference on World Wide Web (WWW) Companion Volume*. Ed. by Jacqueline Bourdeau et al. 2016, p. 717. doi: 10.1145/2872518.2890092. url: http://doi.acm.org/10.1145/ 2872518.2890092; Chengcheng Shao et al. "Anatomy of an online misinformation network," In: *PLoS ONE* 13.4 (2018), e0196087. doi: 10.1371/journal.pone.0196087. url: https://doi.org/10.1371/journal.pone.0196087.

[7] Tim Tangherlini has explored the promise of these middle-sized texts both as digitized fieldwork collections (Timothy R. Tangherlini, „The Folklore Macroscope," Western Folklore 72 (1), (2013): 7-27) and as born-digital community forum posts (TR Tangherlini, V Roychowdhury, B Glenn, CM Crespi, R Bandari, A Wadia, M Falahi, E Ebrahimzadeh, and R Bastani, 2016. "Mommy Blogs" and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites. JMIR Public Health Surveill 2: e166.).

[8] We borrow the term *mesoscale* from manufacturing, where it designates objects between those produced by the manufacturing as many of us imagine it, things like washing machines or cars, which are considered macroscale, and the smaller kinds of objects like the circuits on a computer chip, considered microscale. We are using mesoscale to designate medium-sized texts, and in doing so we recognize that this middle ground has been the focus of consideration within the philosophy of science (see, for example, (Crawford L. Elder. "On the Reality of Medium- Sized Objects". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 83.2 [1996], 191-211.)

[9] The exception to this situation is The Danish Macroscope (Timothy R Tangherlini. "The Folklore Macroscope". In: *Western Folklore* 72.1 [2013], 7-27; Timothy R Tangherlini and Peter Broad-

them, so that we could at least try to see what role it might play in their formation, we decided to create a corpus of transcribed texts that have at least some of their origin in an oral performance and those oral performances could be, if needed or desired, accessed as well. While Youtube offers a wide-rage of such texts with their automated closed captioning processes, such transcriptions are not without their problems and hand editing is often recommended.[10] In response to these criticisms, we sought to work with texts whose transcriptions had been proofed and revised by human eyes. The ever-growing collection of TED talks presented itself as one place to start.

Thus, TED talks offered us a body of middle-sized texts which were already well established in both the public eye as well as in some scholarly considerations.[11] As researchers focused on appraising current mathematical models of textuality and exploring novel models of our own, we think there is more to map than even what Moretti suggests in his declaration that there is an "uncharted expanse of literature … which calls for a maximum methodological boldness."[12] If there is to be a compelling sociology of literature, the boldness is in both examining what we conventionally consider as literature as well as bracketing it to allow into our consideration the wide variety of vernacular literatures which came before, continue to exist alongside, and, even now, morph into ever new forms.

2.2 TED Talks as Talk

One possible objection to a consideration of TED talks as cultural productions is their obviously highly produced and commodified nature, to which one response might be: how is this different from the production of literary forms like the novel, subject to editorial interventions and concerns about publishability based on numbers of readers? Another objection to TED talks might allow them in as culture but not allow them as literary. Two responses to such an objection are possible, one inclusionary and one exclusionary in nature. The inclusive response

well, *The Danish Folklore Macro- scope: Modeling Complexity in the Evald Tang Kristensen Collection*. http://etkspace.scandinavian.ucla.edu/ macroscope.html. [Visited on 11/17/2018]).

[10] National Center on Disability and Access to Education. http://ncdae.org/resources/cheatsheets/ youtube.php(Visited on 01/19/2019).

[11] Wesley Shumar. " 'Being TED': The University Intellectual as Globalised Neoliberal Consumer Self". In: *Learning and Teaching-the International Journal of Higher Education in the Social Sciences* 9.2 (2016), 89-108. doi: 10 . 3167 / latiss . 2016 . 090205; Ana Belen Martinez Garcia. "TED Talks as Life Writing: Online and Offline Activism". In: *Life Writing* 15.4 (2018), 487-503. doi: 10 . 1080 / 14484528 . 2017 . 1405317; George Veletsianos et al. "Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments". In: *PLoS ONE* 13.6 (2018). doi: 10.1371/journal.pone.0197331.

[12] Franco Moretti, "The Slaughterhouse of Literature". In: *Language Quarterly* 61.1 (2000), 207-227.

would argue for getting more humans into not only the archeological-historical record but also the literary record, making literature more important and more robust simply by the depth and diversity of the forms within which individuals can express themselves and their moment in time. The exclusive response argues that literature, as a form of production, must be bounded, and the only way to begin to discern those boundaries is to know what literature is not. If TED talks are not literary, then how are they not and what questions does that raise about past and current conceptions of literature or what possibilities does it offer the domains of literary production and the study of that production in terms of useful ideas, structures, or methods? We think by being hybrid texts, TED talks offer investigators an opportunity to examine where talk becomes marked as a particular kind of performance, a subject that has been a consideration of folkloristics for quite some time.

That noted, the data being released here are the transcripts of TED talks, which are not, as folklorists and linguistic anthropologists would be keen to note, not the talks themselves. Such transcripts offer an hybrid object for analysis of cultural phenomenon, especially with the additional data about the speakers, the view counts, and the discussion threads attached to each talk. The widespread popularity of TED talks means there is a broad sense of what a TED talk is in terms of form, content, and style. And their length places them firmly within the realm of the kinds of discursive production in which most human beings engage.

This bridge to the rest of humanity is not without its difficulties. While the transcripts are not the talks themselves, they are a fair representation of an oral performance, something with which folklorists and anthropologists have long been concerned. While TED talks are obviously the result of practice and polish, it is entirely possible to argue that many conversational texts we encounter in daily life are themselves the net result of similar amounts of practice and polish either in the mouth of the person performing them or in the mouths of previous performers who have passed the text along. [13]

0:14 Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful. I have been blown away by

[13] The notion that a given text, be it a story or song, gets refined through its re-tellings is central to a number of folkloristic theories. Chronicling the people and ideas of late twentieth century American folklore studies, Bill Ivey notes that "Unlike fixed texts of civilizations, the lore of the folk lives in oral tradition, and the subtle changes introduced through face-to-face transmission gives folklore texts special validity, a truth fored through shared expression refined over time" Bill Ivey, *Rebuilding an Enlightened World: Folklorizing America,* Indiana University Press, 2018, 49. Describing the effects on one of the more prosaic folklore forms, the legend, Bill Ellis notes that "the more narrators tell their stories, the more they will focus on the narrative elements that most efficiently re-present the most successful translation of the events described" Bill Ellis, *Aliens, Ghosts, and Cults: Legends We Live,* University Press of Mississippi, 2001, 64.

this conference, and I want to thank all of you for the many nice comments about what I had to say the other night. And I say that sincerely, partly because (Mock sob) I need that.
0:40 (Laughter)
0:45 Put yourselves in my position.
0:47 (Laughter)
0:54 I flew on Air Force Two for eight years.
0:57 (Laughter)
0:59 Now I have to take off my shoes or boots to get on an airplane!
1:02 (Laughter)
1:05 (Applause)

Table 1: Table 1: An excerpt from a Al Gore's TED talk in 2006 with parentheticals from both the speaker and the audience included in the transcript text along.[14]

So, while TED talks are far from conversational discourse, they offer researchers the opportunity to explore what dimensions of oral textuality might be open to algorithmic analyses. TED talks, qua texts, represent an interesting oral-literary hybrid: they are performed orally but not spontaneously. They are clearly the product of considerable scripting, which may involve writing but may simply involve a lot of rehearsing. The talks themselves are given without reference to a script, unlike other performances (e.g., political speeches) but are more scripted than the kind of daily performances which have been the focus of folkloristic and anthropological inquiry. Most importantly, the transcripts which are a part of the data set are not, however, the product of the speakers but of the TED organization: they are produced after the fact, by a combination of TED staff and crowd sourcing. A TED representative we contacted reported the following:

> [A] ll transcriptions for TED talks are created by our Translation team, aided by hundreds of contributors in our "TED Translator" program. TED Translators are volunteers from around the world who enable the inspiring ideas in them to crisscross languages and borders. They use a free, online subtitling tool called *Amara*, which volunteers can use to find talks, create subtitles, peer review transcriptions and trans- lations, and message other TED Translators.[15]

[14] The notion that a given text, be it a story or song, gets refined through its re-tellings is central to a number of folkloristic theories. Chronicling the people and ideas of late twentieth century American folklore studies, Bill Ivey notes that "Unlike fixed texts of civilizations, the lore of the folk lives in oral tradition, and the subtle changes introduced through face-to-face transmission gives folklore texts special validity, a truth fored through shared ex- pression refined over time" Bill Ivey. *Rebuilding an Enlightened World: Folklorizing America*. Indiana University Press, 2018, p. 49. Describing the effects on one of the more prosaic folklore forms, the legend, Bill Ellis notes that "the more narrators tell their stories, the more they will focus on the narrative elements that most efficiently re-present the most successful translation of the events described" Bill Ellis. *Aliens, Ghosts, and Cults: Legends We Live*. University Press of Mississippi, 2001, p. 64.

[15] Per correspondence with TED Support. See also the page explaining the transcription process as part of the larger translation process: https://www.ted.com/participate/translate/transcribe.

While the relationship between any particular in-house transcription effort and the crowd-sourced effort is not entirely clear, there is a qualitative difference between these transcripts and transcripts available via automated systems such as those found on YouTube or a script from which a speaker worked. The result of the TED in-house transcription efforts are transcripts that look like the beginning of Al Gore's 2006 talk excerpted in Table 1.

The transcripts are formatted as tables on TED web pages, with approximate time codes on the left and blocks of text on the right. Also on the right are parentheticals from both the speaker's performance as well as the audience's reaction. In the example, taken from Al Gore's talk in 2006, we have both an internal parenthetical, describing a particular moment in Gore's presentation, "(Mock sob)", as well as several examples of the audience's response to Gore, "(Laughter)". None of these are a part of the text itself, however, and when compiling vocabulary and modeling topics, we have removed these. (But we did want to be able to keep them in mind when, for example, analyzing the texts for sentiment.)

Because the texts available on the TED website are transcripts of a spoken performance, we had to confront at least one question about the status of these texts: accuracy. That is, how closely does a transcript track with an author's intention? Accuracy here is in two dimensions: words and syntax. That is, does a given transcript accurately capture the words the speaker used and the way in which the speaker strung them together? We chose a number of presentations to watch with the transcripts next to us and we were satisfied that, especially when it came to matters of lexicon, the transcriptions were of a high quality. Whether a speaker meant a semi-colon, a colon, or a period was not something we felt we could gauge: so how a speaker intended coordination of clauses and how those clauses are represented in a transcript is not something that can be relied upon. Researchers interested in syntactical dimensions will need to be aware of this.

Setting aside the matter of transcription, TED talks present the possibility of exploring the nature of performance. Over the last fifty years, folklorists, anthropologists, linguists and communications scholars have built upon an initial delineation of performance as one way to think about verbal art, expanding it to address a wide range of human behavior associated and/or bound up with verbal acts.[16] Such notions of performance extend from formally-staged events, like

[16] Folklorist Richard Bauman, pursuing a line of inquiry begun by linguist Dell Hymes(Edwin Ardener, ed. *Soci- olinguistics and the ethnography of speaking*. London: Tavistock Press, 1971, pp. 47-93), was the first to articulate the notion of "verbal art as performance"(Richard Bauman. "Verbal Art as Performance". In: *American An- thropologist* 77.2 [1975], 290-311). Subsequent work by other scholars and scientists elaborated on the nature of performance and its possible boundaries (Elizabeth

those on TED talks that also include staging for the screen, to a wide variety of informal events that include whispered gossip in high school corridors. Beyond the fundamentals of average speaking rate that one can deduce from the metadata already available in the current data set, further explorations of prosody are possible that might begin to establish how TED talks participate in both fundamental forms of prosody and/or, as parodies suggest, innovate in particular (and peculiar) ways loudness, and rhythm, timing and speaking rate.[17] A macroscopic examination of intonation patterns and pitch, speech stress, speaking rate, and rhythm and other paralinguistic and extralinguistic features promises to help us understand how gender, domain expertise, and rhetorical situations play out.

2.3 TED Talks as Data

The current output of this project, and the materials we are releasing, are collated versions of the TED talks with meta information about both the talks and the speakers themselves. Below we detail the central data sets contained within this release as well as some of the supplementary materials available. The data can be found at our GitHub page: https://github.com/kinnaird-laudun/data/tree/master/Release_v0. The data itself is comprised of information from 2656 talks downloaded from the main TED site in the summer of 2018. We used the list of talks maintained as a Google Sheet that Dan Colman reported on *Open Culture*.[18] Our method for downloading and cleaning the talks into a series of CSVs is explained in the next section.

While the TED site publishes transcripts to talks at TEDx events, we are narrowing our focus to those at the main TED event and the seven major series. These satellite events are organized locally and use their own criteria for inviting speakers. This means that the talks at the main TED events and the TEDx events have different kinds of selection bias associated to them. Furthermore, not all TEDx talks are published on the TED website, which introduces another layer of selection bias. So as not to give the impression that the dataset includes all (or even a

Fine. *The Folklore Text: From Performance to Print*. Indiana University Press, 1984).

[17] Some early work on performances by poets by MacArthur, Zellou, and Miller focused on 12 prosodic features and suggests that, at least for formal readings of poetry, there are "four stylistic tendencies that begin to characterize reading styles more precisely than the binary of neutral vs. expressive"(Marit MacArthur, Georgia Zellou, and Lee Miller. "Beyond Poet Voice: Sampling the (Non-) Performance Styles of 100 American Poets". In: *Journal of Cultural Analytics* [2018]).

[18] https://docs.google.com/spreadsheets/d/1Yv_9nDl4ocIZR0GXU3OZuBaXxER1blfwR_XHvklPpEM/; Dan Colman. *1756 TED Talks Listed in a Neat Spreadsheet*. http://www.openculture.com/2014/06/1756- ted-talks-listed-in-a-neat-spreadsheet.html. 2014.

representative sample of) TEDx talks, we have chosen not to include any of the TEDx talks presented on the TED website in this release.[19]

In the data release directory are the following five files:

- a CSV exported from the Google sheet[20] in May 2018
- a CSV containing the metadata as well as the transcripts of the talks that occurred at the annual TED event: TEDonly_final.csv
- a similar CSV as the one above but of the talks from one of the seven major series (TEDActive, TEDGlobal, TEDMED, TEDSummit, TEDWomen, TEDYouth): TEDplus_final.csv
- TEDonly_speakers_final.csv - CSV of the talks at the main TED event collated with meta information about the speakers
- TEDplus_speakers_final.csv - CSV of the talks at one of the seven major series collated with meta information about the speakers
- a copy of the TED license (CC BY-NC-ND) at the time of this release, the same license under which this release is made.

Both the TEDonly and TEDplus files in this release include the following features for each talk:

- Official TED identification number
- Public URL
- Name(s) of the speaker(s)
- Headline assigned by TED
- Description of the talk as it appeared on the talk's main page

- Particular TED event at which the talk occurred
- Duration of the talk (in seconds)
- Date on which the talk was published on the TED site
- Tags assigned to the talk by TED
- Number of views the talk had received as of summer 2018
- Transcribed text of the talk

The two CSV files with the meta information provided on the TED website about the speakers include the following, delineated by speaker:

- the speaker's occupation,

[19] The larger repository of which this data release is a part includes a working directory one level above the release with our Jupyter Notebooks that can be adjusted to process the existing TEDx talks into a collated dataset.

[20] https://docs.google.com/spreadsheets/d/1Yv_9nDl4ocIZR0GXU3OZuBaXxER1blfwR_XHvklPpEM/.

- the speaker's introduction as given on the TED site, and
- the profile of the speaker beginning with the phrase "Why you should listen to …" In addition to these CSVs, we also are making available:
- The URL lists which we derived from the original Google sheet and used as input for wget.[21]

Originally, we wished to include all the HTML files involved for all of the talks, including the HTML for the speaker pages, the page in which a talk is embedded, the transcript pages, and, finally, the discussion page for each talk, but given the generosity of the TED organization in allowing us to share the data above in this structured fashion, we have elected to set aside that data from release. For those interested in exploring the changing nature of the organization's web infrastructure, which has included some substantive changes to the HTML, we are happy to make the data available upon request. These are the complete HTML, available to be parsed in various ways.[22]

The Jupyter notebooks contain our Python work as well as a log of our command line operations. While we recognize that both Python and Jupyter are actively developed and maintained, the notebooks here are mostly included as documentation of our process. In keeping with recent discussions about reproducibility of research, the notebooks make it possible for interested individuals to see what we do and to reproduce those results however they see fit.[23]

There are additional files in the GitHub repo in the parent directory of the data. These files are described in the associated Jupyter Notebooks within the release directory associated with this paper.[24]

The release of this data falls within the scope of the Creative Commons license that TED operates under—CC BY-NC-ND 4.0—as this is an analytical, collated, and clean dataset of the TED talk transcripts. Additionally, we are sharing the transcripts in a new format without creating a derivative of them, which is explicitly permitted under the CC BY-NC-ND 4.0 license. Additionally, we have

[21] For further considerations of reproducibility within cultural analytics see (Andrew Piper. *Do we know what we are doing?* 2019. url: http://culturalanalytics.org/2019/04/do-we-know-what-we-are-doing/

[22] We offer the HTML files of the TED talks and the speaker profiles both in keeping with the permissions of the TED organization and as an archive of these sources in their current form. As we note below, some of our original explorations in 2016 used different code to parse the pages, revealing that the TED site has undergone at least one revision since then. The Google sheet itself has also seen the addition of columns and a change in name of some of the headers.

[23] We cannot guarantee that the code will itself run as is as Python and Jupyter evolve nor that the code will work out of the box on any particular machine setup.

[24] The GitHub directory is located at: https://github.com/kinnaird-laudun/data.

sought and gained explicit approval from TED for this release.[25]

3 Methods for Acquiring and Cleaning the Data

Due to the storage methods on the TED website and the evolution of the pages over time, collating information about the TED talks is a non-trivial process. Beginning with the Google sheet[26] listing the URLs of the talks, the process involves a number of steps: (1) acquiring the HTML files, (2) extracting the desired meta-information from within the hybrid JSON/HTML structure of a file, (3) parsing the speaker information from the separate set of files and joining it to the appropriate talks, and (4) inspecting all the output files by hand for errors and missing data and then making those corrections and additions. As the enumeration of the process makes clear, the files within our data release build on each other. The exported Google sheet is leveraged to create the TEDonly and TEDplus files by cataloging the locations of those talk pages. The files with the speakers' meta information are expansions of the files without.[27]

3.1 Acquisition of TED transcripts

Like many websites serving as media delivery platforms, the TED site continues to evolve as the collection of TED talks grows with each event. This evolution can be viewed through the growth of the Google sheet[28] that catalogues the talks on the TED site. When Open Culture first reported the existence of the sheet in 2014, the list contained 1756 TED talks.[29] In 2016, when we first decided to begin collaborating and made TED talks a place to start having a dialogue about the application of statistical methods to humanities inquiries, the list contained 2209 talks. At the time of our update in June 2018, the list had grown to 2755 talks.

[25] Specifically, TED has granted us permission for "this use on a one-time, non-precedent setting basis." A copy of the email, which serves as the official license, can be provided upon request.

[26] https://docs.google.com/spreadsheets/d/1Yv_9nDl4ocIZR0GXU3OZuBaXxER1blfwR_XHvklPpEM/.

[27] Please note that the Jupyter notebooks in the data directory (one level up from the release) should be considered historical in nature: while the code in them can be run, in some cases input and output files have subsequently been moved and/or renamed as our own processes and priorities became clearer to us. We welcome their use: just know that some editing of the loading and saving of the working files may be needed.

[28] https://docs.google.com/spreadsheets/d/1Yv_9nDl4ocIZR0GXU3OZuBaXxER1blfwR_XHvklPpEM/.

[29] Colman, *1756 TED Talks Listed in a Neat Spreadsheet*, op. cit.

The number of talks can be partially explained by the inclusion of talks at TEDx events that have become very popular. TEDx events are organized locally but without explicit oversight from the TED organization: "TEDx events are planned and coordinated independently, on a community- by-community basis, under a free license from TED".[30] As the selection criteria is not consistent between TEDx events and the main TED events, including their seven major series - TEDActive, TEDGlobal, TEDMED, TEDSummit, TEDWomen, TEDYouth - and because not every TEDx talk is included on the main TED website, we narrowed our data collation to only the main TED events and series: so the speaker files are only applicable to the main and major series. (All the talks available for download in June 2018 are available in the merge file.) The Google sheet[31] serves as the starting point for our data collation, as it lists the main URL for each TED talk. Using this list, we download the main pages and then amend the URLs to download the transcripts and discussions by appending `"/transcript"` and `"/discussion"`. All of the HTML pages are part of the supplemental materials to this data release, both as a source of additional information about the TED talks not currently part of the CSVs in this data release and as an archive of the TED website itself.[32] Also part of this data release is the code from the most recent parsing of the files (June 2018). We note that between 2016 and 2018, the TED website was revised such that much of the code written for parsing the HTML pages in 2016 had to be rewritten in order to work on the files downloaded in 2018.[33]

3.2 Extracting Descriptions and Transcripts

From the collection of pages, we then separate the meta-information about the talks from the syntax of the website. The TED webpages are a combination of conventional HTML and a hybrid JSON, meaning that this separation step is an iterative process. Using Python's BeautifulSoup and json modules as well as some regular expressions, we extract a tuple of the talk information–including a talk's identifier, description, view count, and the TED event at which it occurred–which

[30] https://www.ted.com/about/programs-initiatives/tedx-program.

[31] https://docs.google.com/spreadsheets/d/1Yv_9nDl4ocIZR0GXU3OZuBaXxER1blfwR_XHvklPpEM/.

[32] While the archival usage is not obvious, the dynamic nature of the TED website was driven home to us when we began work on the downloaded files in 2018 and soon discovered how much our code from 2016 had to be revised or rewritten to work in 2018. And, at least in the case of accompanying time signatures throughout a talk, some of the data was no longer available.

[33] The 2016 files and the accompanying code are available by request to anyone interested in exploring differences between both the code and the content of the TED website itself.

we then placed into a row in a CSV file.[34]

Getting the text of the talk was a similar process, but due to the coding of these files, this step involved collecting all the divs of the class`Grid cell flx-s:1 p-r:4.`As shown in Table 1, the transcripts provided by TED's in-house transcription efforts include more than just the words spoken by the speaker(s). The transcripts are formatted as tables on TED web pages, with approximate time codes on the left and both blocks of text as well as parentheticals external to the text and, in most cases, to the speaker on the right.[35] So for example, Al Gore's talk from Table 1 would start as:

> Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful. I have been blown away by this conference, and I want to thank all of you for the many nice comments about what I had to say the other night. And I say that sincerely, partly because (Mock sob) I need that. (Laughter) Put yourselves in my position. (Laughter) I flew on Air Force Two for eight years. (Laughter) Now I have to take off my shoes or boots to get on an airplane! (Laughter) (Applause) …

In the above example, we have both an internal parenthetical, describing a particular moment in Gore's presentation which is part of the performance, the mock sob, as well as several examples of the audience's response to Gore, laughter and applause. We leave these parentheticals in with the understanding that different analysts will have different understandings of what is, and is not, part of the text. In our work, we remove the parentheticals for the operations that focus on lexicon and include them when dealing with affect.[36]

3.3 Adding Speaker Information

TED talks often gain popularity due to the fame of the speaker, but there are also cases of speakers rising to the national conversation due to a viral TED talk. The close relationship between TED talk transcript and speaker motivates the need for collated data files with meta information for both the speaker(s) and the talk

[34] A detailed description of the process is available in the "Parsing the Descriptions" Jupyter notebook that is included with this data release.

[35] See the "Parsing the Transcriptions" Jupyter notebook in the data release.

[36] There are additional notes on parentheticals, as well as the regex we found useful, in the Jupyter notebook on tokenization. We emphasize that users of the data will need to process the text as they think best for their purposes.

itself. This process begins with the gathering of speaker information, parsing of talks' speakers, and matching speaker meta information.

The extracting of the speaker information–their occupation, their introduction, and their profile–is a similar process to the extraction of meta information about the talks. Each speaker has a page within the speaker index page, e.g., `/speakers/ellen_t_hoen`. After downloading all the speaker pages using the parsed relative URL for each speaker, we created another text file to feed wget and extract the various types of meta information.

Demographic information such as race, gender, and age of the speaker is not provided by TED on the speaker pages. This kind of demographic information would need to be gleaned from third-party sources. However, initial work has been done to automatically detect each speaker's self-identified gender from their introduction and profile text leveraging the proportion of gendered pronouns in these short texts. This work is in early stages and is beyond the scope of this paper that is a collation of information tracked and stored by TED.

While not all talks have more than one speaker, for the talks that do, we separate individual speakers into their own columns. This step is a simple splitting of the speakers string on the word 'and.' This separation process can occur in parallel to the extraction of speaker information. This step is crucial because speaker meta information is usually by individual speaker and not by groups of speakers. For example, while the two co-founders of Google Sergei Brin and Larry Page have given TED talks together, they each have their own profiles within the TED eco-system. Once the speakers are separated into their own columns, the adding of the majority of the speakers' meta information uses operations like merge and join from pandas in Python.[37]

3.4 Hand-coding

We would be remiss not to note that in collating information about TED talks into one data set, small amounts of hand-coding are necessary. The reasons for each of these hand-coding steps vary, but each instance of hand-coding is done following visual inspection of the data and confirming that (1) hand-coding is necessary and (2) finding the missing information within the TED eco-system. Below are a few examples with their Talk ID numbers.[38]

[37] Details of this merging can be found in the associated Jupyter notebooks.

[38] The transcripts that we are releasing as part of this data release are labeled as transcript.##. The numbers at the end of the file name are not the talk identification number as given by TED. The talk identification number is instead buried within the HTML coding.

- *Bill and Melinda Gates* as speakers - Talk ID 1964: This is one example that fails due to the speaker separation process. In these TED talk HTML files, the speakers are listed at ‘Bill and Melinda Gates.’ This means that when we split on ‘and’, we end up with ‘Bill’ as speaker one and ‘Melinda Gates’ as speaker two. For colloquial usage–like a website–stating the speakers of this talk as ‘Bill and Melinda Gates’ makes absolute sense, but as speaker entities, they are ‘Bill Gates’ and ‘Melinda Gates.’

- *Gustavo Dudamel and the Teresa Carreño Youth Orchestra* - Talk ID 466: This talk also fails due to the speaker separation process but for a slightly different reason. Here the presenter is both the conductor - Gustavo Dudamel - and his orchestra - the Teresa Carreño Youth Orchestra - as one entity. This is a more formal use of ‘and’ than in the above example. By splitting on ‘and’ we incorrectly split a speaker into two parts that are not speakers within the TED eco-system in their own right. We find a similar issue with Tals ID 1972, as Gabby Giffords and Mark Kelly are treated as one speaker entity in the TED ecosystem.

- *Talks with missing speakers* - Talk IDs 1722 and 1786: As we inspect the data sets after adding the speaker information to each talk, we noticed two talks that had no speakers. Investigating further on the TED website, these talks did have speakers but where that information was stored on those talks’ pages was coding differently than other pages.

- *Duplicate speaker names* - TED IDs 72, 211, 1785, 3633, and 8420: Within our listing of TED speakers, we have two sets of duplicate names. We have two speakers named Chris Anderson (one the curator of TED and the second a drone maker) and two named Michael Green (one an economist and the other an architect). So when we join speakers with their talks, the python code does not know which Chris Anderson (or Michael Green) to select, so it creates a talk row for each of the identically named speakers. For the hand-coding, we found these duplicate talks and selected the correct speaker.

Each of the above examples is unique in why they fail to correctly match the speaker meta informa- tion to the talk. Furthermore, writing one single script to deal with each of these cases correctly, while possible, would be time better served by doing manual adjustments as the cases are (1) very limited in number and (2) rely on our contextual understanding. The first two examples stumble on colloquial and formal uses of *and*. Writing a script to check for colloquial versus formal uses of *and* would require a computer to understand the differences in these usages. The third example would require a broader scraping of the HTML files which would likely mean that we would return additional information that is not relevant to our search, which in turn would require additional cleaning. The fourth case requires one to understand the content of the talk and the occupations of the potential speakers in order to correctly match the right person with the duplicated name to each talk. This additional check in code requires either a lengthy list of possibilities to be checked or a sophisticated artificial intelligence system.

We also run into unicode issues when matching speaker's meta-information with their talks. For some speakers with accents in their names, the CSVs with the talk information did not store their names correctly, either adding strange subsequences of characters or changing the letters in the name.[39] This of course results in speaker meta-information not being matched correctly. To resolve this, we do a version of the procedure resolving the duplicate speaker names above. In our resolution, we also overwrite the incorrect encoding of speakers' names in the names columns.[40]

The hand-coding step highlights the necessity for careful collation of data and for human oversight during that process. In working with the TED talk and speaker meta-information, we find examples of incorrect spellings due to errant encoding, conflicts between colloquial and intentional use of 'and' for denoting multiple or single speakers, and matching talks with speakers when two speakers share the same name. The diversity of these issues and the kinds of resolutions that they require illustrate that humans need to be an engaged part of the data wrangling process.

3.5 Preview of the Cleaned TED Data

In our data presentation of the TED talk transcripts, each talk is its own row with the different types of meta-information listed in the columns. Table2is a snippet of one talk's row in the version that includes meta-information of the speaker(s): we have transposed the rows and columns for the sake of space in order to provide readers with an example of what is available for any given talk.

[39] It is interesting that it seemed to only be the files with the talks that have issues with accents while the files containing only speaker meta information encode accents within names as expected.

[40] To ensure that our hand coding decisions and work are also in the dataset without speaker metadata, we create our final version of that data by simply removing the columns associated to speaker metadata from the combined talk and speaker file. The only columns left are those associated to the talk meta-information and those that list the name(s) of the speaker(s).

Talk_ID	1
public_url	https://www.ted.com/talks/al_gore_on_averting_climate_crisis
headline	Averting the climate crisis
description	With the same humor and humanity he exuded in ”An Inconvenient Truth,” Al Gore spells out 15 ways that individuals can address climate change immediately, from buying a hybrid to inventing a new, hotter brand name for global warming.
event	TED2006
duration	00:16:17
published	6/27/06
tags	alternative energy, cars, global issues, climate change, environment, science, culture, sustainability, technology
views	3266733
text	Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful. I have been blown away by this conference, and I want to thank all of you for the many nice comments about what I had to say the other night. And I say that sincerely, partly because (Mock sob) I need that. (Laughter) Put yourselves in my position. (Laughter) I flew on Air Force Two for eight years. (Laughter) Now I have to take off my shoes or boots to get on an airplane! (Laughter) (Applause) …
speaker_1	Al Gore
speaker1_occupation	Climate advocate
speaker1_introduction	Nobel Laureate Al Gore focused the world's attention on the global climate crisis. Now he's showing us how we're moving towards real solutions.
speaker1_profile	Why you should listen: Former Vice President Al Gore is co-founder and chairman of Generation Investment Management. While he's is a senior partner at Kleiner Perkins Caufield & Byers, and a member of Apple, Inc.'s board of directors, Gore spends the majority of his time as chair of The Climate Reality Project, . . .
speaker_2	
speaker2_occupation	
speaker2_introduction	
speaker2_profile	
speaker_3	
speaker3_occupation	
speaker3_introduction	
speaker3_profile	
speaker_4	
speaker4_occupation	
speaker4_introduction	
speaker4_profile	

Table 2. A transposed row from TEDonly_speakers_final.csv of data showing labels in use and the associated data: please note that the talk itself and the speaker profile have been truncated for the sake of space.

4 Conclusions and Future Work

The main contribution of the current work are collated data sets of TED talks from both an inclusive set of all talks published on the TED website as well a more focused set of the main TED events sponsored by the TED organization itself. These collections, whether of the few thousand inclusive set or the 1754 talks with extended speaker information, offer well-documented texts. Their rich metadata offer researchers in the humanities and data science a number of avenues for considering texts in various contexts, be they domain, gender, popularity, or the discourse that accompanies them. While one version of this data focuses solely on meta-information about the talk itself, a second version also includes meta-information about the speakers. These data sets have been contextualized through a tri-lens discussion of TED talks as texts, talk, and as data.

There is, of course, always more to do when it comes to data processing and munging. For example, compiling the additional data for the many (though incomplete on the TED website list of) TEDx talks that would, perhaps, give further insights into how genders and disciplines, as well as topics, fare outside of central TED events. To this end, in this work, we have articulated our data collation methods and made available the code used during this process.

One of the immediate outcomes of doing this work is how the process itself asks questions and forces analysts to make decisions which must then be explained and defended: for example, determining what constitutes a talk. Our own work on TED talks is fundamentally focused on making the text of the talks the center of much of our analysis: what words are used by what kinds of speakers, as well as how those words constitute larger semantic formations, like topics, or reveal rhetorical moves or changes in modality that would allow us to infer, if anything, the relationship between texts and their non-textual traits (like views)? As a result, talks which featured texts with no to few words were less interesting to us, despite these largely being musical performances. (The possibility of asking questions about the nature of talks in which talk does not feature is something we leave for another time.)[41]

In addition to their middle-sized nature as well as their hybrid status as oral-literary texts, TED talks also come with a great deal of metadata. The ability to compare the two lengths of a text, the time it takes to be delivered versus the number of words within it, is but one possibility among many. While popularity is one gauge for a talk's importance, we think the increasing interest in methods for detecting text re-use could reveal which talks have had more subtle, less easily discerned affects on contemporary discourses. With all the data available in the TED talk data set, we hope to encourage new lines of research for investigating cultural phenomena.

In the introduction to their exploration of possible intersections between cultural analytics and critical race studies, So, Long, and Zhu note that cultural analytics, with the cultivation of large data sets of texts and the adoption of computational technologies, has pursued "an expanding array of topics, including genre and

[41] Richard Bauman. *Let Your Words be Few: Symbolism of Speaking and Silence Among Seventeenth-Century Quakers.* Waveland Press, 1983.

cultural prestige".[42] Cast often as thought leaders and the shapers of public conversations on topics, TED speakers and their talks offer analysts an opportunity to examine the relationship between ideas and the marketplace(s) in which they are embedded. With the variety of speakers addressing diverse topics in a highly prestigious forum, the talks - along with their transcripts and meta-data - provide insight into current practices in the commodification, and to some degree the reception, of ideas, images, and events.[43] With a plethora of metadata not only about a talk's performance date, its publication date, its popularity, and a rich set of data about its speaker(s), the current data set opens new avenues of research and is just a first step towards understanding the intersection of texts and the world they hope to shape and are, in turn, shaped by.

[42] Richard Jean So, Hoyt Long, and Yuancheng Zhu. "Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000". In: *Journal of Cultural Analytics* (2019). doi: 10.22148/16.031.

[43] We note that more work should be done to understand the bias that is present among the TED talks given that the TED speakers as a group do not match the racial, gender, or socioeconomic diversity of the general population in the United States or globally.