Peer-reviewed version available at *Data* **2019**, *4*, 48; doi:10.3390/data4020048

Article

From a Smoking Gun to Spent Fuel: Principled Subsampling Methods for Building Big Language Data Corpora from Monitor Corpora

Jacqueline Hettel Tidwell [1,*]

[1] Social Energy Atlas, Franklin College of Arts and Sciences, University of Georgia; jacqueline.tidwell@uga.edu

Abstract: With the influence of Big Data culture on qualitative data collection, acquisition, and processing, it is becoming increasingly important that social scientists understand the complexity underlying data collection and the resulting models and analyses. Systematic approaches for creating computationally tractable models need to be employed in order to create representative, specialized reference corpora subsampled from Big Language Data sources. Even more importantly, any such method must be tested and vetted for its reproducibility and consistency in generating a representative model of a particular population in question. This article considers and tests one such method for Big Language Data downsampling of digitally-accessible language data to determine both how to operationalize this form of corpus model creation, as well as testing whether the method is reproducible. Using the U.S. Nuclear Regulatory Commission's public documentation database as a test source, the sampling method's procedure was evaluated to assess variation in the rate of which documents were deemed fit for inclusion or exclusion from the corpus across four iterations. The findings of this study indicate that such a principled sampling method is viable, thus necessitating the need for an approach for creating language-based models that account for extralinguistic factors and linguistic characteristics of documents.

Keywords: corpus linguistics; language modeling; big data; language data; databases; monitor corpora; documentary analysis; nuclear power; government regulation; tobacco documents

1. Introduction

We now exist in the Age of Big Data [1]. Regardless of one's discipline or area of interest when it comes to language, the influence of Big Data culture on the analysis of language is undeniable. Computing technology that can handle increasingly large amounts of data continues to emerge. The increase in focus on the computational analysis of large collections of text was seen in the field of linguistics even before our entering into the Age of Big Data and supercomputing technologies. A study conducted in 1991, reports that from 1976 to then, the number of corpus linguistic studies doubled for every five years [2,3]. One of the primary reasons why this increase occurred is due to the introduction of personal computers to the technology marketplace [4], as they facilitated the ability to create text-based models that were explicit, consistent, and representative of the population they signified. In much the same way that the personal computer precipitated an increase in corpus-based studies, our ability to access vast numbers of readily available machine-readable language resources and storage capabilities for creating high volume corpora has changed the shape of language-based modeling methods.

1.1. The Rise of Big Language Data

Big Data not only refers to large data but more importantly to diverse and complex data that are difficult to process and analyze using traditional methods. Big Data is notable because of its relationality with other data and networked nature [6,15]. Big Language Data corpora are not merely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2018 doi:10.20944/preprints201812.0016.v1

Peer-reviewed version available at *Data* **2019**, *4*, 48; doi:10.3390/data4020048

larger corpora; they are highly relational models that have the potential for providing insights into why variation occurs in different contexts. Creating the largest collections of machine-readable language does not necessarily mean better analysis and more robust levels of understanding. Some of the most massive available corpora, for example, the Time magazine corpus [7] and even the Web [8], have not been compiled using rigorous, systematic protocols and may very well provide a biased perspective on language in use [9]. Addressing these metadata characteristics of sampled corpora in a statistically rigorous way is of significant concern if the goal is to investigate variation in the transmission or reception of concepts communicated in written or spoken language.

Despite these severe challenges to contemporary research involving qualitative language data, corpus design methodologies are an understudied component of investigations into the use and variation of English in specific digital contexts [10]. With the influence of Big Data culture on qualitative data collection, acquisition, and processing, it is becoming increasingly important that social scientists begin endeavoring to understand why the ways in which they collect data affect their resulting analyses. For example, In the case of monitor corpora, the Web, and even databases that are regularly having content added to them, their inherently dynamic nature typically renders them unsuitable for comparative studies since one cannot perform descriptive linguistic analysis on them: they are continually changing [11]. It is not the goal of this article to advocate for throwing the baby out with the bathwater regarding dynamic and unsampled Big Language Datasets. Instead, the objective is to demonstrate a method for leveraging existing Big Language Data of this nature and transforming them into Big Language Data corpora that adequately model and reflect the purpose of the analysis.

1.2. Sampling Parameters for Big Language Data Studies

All types of Big Data, whether they are language based or not, are by definition unwieldy and difficult to make sense of without the use of methods for making them more manageable. The easiest way to work with Big Data is actually to avoid it by subsampling [12]. Corpus Linguistics is one such method for creating subsets of Big Language Data through the systematic collection of naturally occurring texts, or "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" [13]. There is a considerable amount of effort and planning that go into the design of corpora that enable us to understand better language as it is really used.

Big Language Data is "Big" because of its highly relational and complex nature. Language is, in fact, a complex system, as defined and studied in physics, evolutionary biology, genetics, and other fields [14]. One of the reasons why analysis of Big Language Data is so provocative is because it facilitates the observation of emerging trends from a complex network of relationships [15]. Emergence is one of the defining characteristics of complex systems, and in language it comes in the form of a non-linear, asymptotic hyperbolic curve, or A-Curve, that has been documented extensively in linguistic survey data of American English from the Linguistic Atlas Projects [14,16,17]. The resulting language used occurs in scale-free networks where the same emerging pattern occurs at every level of scale for linguistic frequencies from small groups of speakers to national ones.

The objective of creating corpora from Big Language Data so to understand the population from both textual and social perspectives at different levels of scale within the complex system is to create distinct subsets of the language employing rigorous sampling principles. One tool for defining specific subsets of language data is through the use of a sampling framework. A sampling framework is essentially a list, map, or other specification of elements or characteristics of a population of interest from which a sample may be selected [18]. Sampling frameworks are of critical importance for creating a subsample of a Big Language Dataset that can be used to scale-up or generalize about the population of interest as a whole. The use of methods based on random sampling that provides every member of the population an equal opportunity to be sampled is quite common in modern sociological survey research: e.g. election polling [19,20]. Employing such an approach affords a linguist the confidence that the corpus is representative of the complex system they are attempting to model.

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

This issue of representativeness is defined within the sampling parameters established before the corpus is created. Two aspects of a population of interest must be defined when creating a traditional sampling framework: a definition of boundaries of the population, or the texts to be included and excluded; and a definition of the hierarchical organization to be included, or what text categories are included [21]. Traditionally in corpus linguistics, both the sampling framework and population of interest are defined by linguistic or text-based characteristics. Linguistic representativeness is dependent on the condition that a corpus should represent the range of text types in a population. The notion of sampling based on characteristics of the people authoring, speaking, or transmitting the language is considered an alternative to sampling frameworks: demographic sampling [22]. In demographic sampling, data for a corpus is selected by person, entity, or agent rather than text. Both of these approaches for subsampling Big Language Data are systematic and allow for the creation of corpora that reflect a specific population of interest for computational analysis.

2. Materials and Methods

When creating a representative, specialized reference corpus subsampled from Big Language Data sources, such as large, dynamic databases of texts or online repositories of documents, it is imperative that a systematic approach for creating a computationally tractable model be employed. Even more importantly, any such method must be tested and vetted for its reproducibility and consistency in generating a representative model of a particular population in question. This article will both consider and test one such method for Big Language Data downsampling of digitally-accessible language data.

2.1. Creating the Tobacco Documents Corpus

In 2004, W.A. Ketzschmar et al. [23] proposed a principled sampling method for creating a reference corpus from a collection of documents from the tobacco industry (TIDs). In the fall of 1998, a settlement was reached by the National Association of Attorneys General and seven major United States tobacco industry corporations in order to impose regulatory measures on the tobacco industry. As a result, the seven corporations were required to release all industry documents to the public that were not considered attorney-client privileged nor to have contained proprietary trade information.

They proposed a two-stage, iterative approach for sampling, with a purposely designed sampling framework based on a well-defined population of interest [24]. The first phase, or pilot corpus, was to be drawn in order to determine how text types should be classified, as well as estimating their proportions within the population of interest. Therefore, special attention needed to be applied to text types for the pilot corpus upon which the reference corpus would be built in order to avoid skewing the data. However, before the Tobacco Documents Corpus (TDC) pilot could even be created to investigate this variety, they had a slight issue from a theoretical standpoint with their sampling population.

In order to deal with large-scale monitor corpora like the Tobacco Documents for comparative corpus-based research, the entire body of documents was sampled according to a fixed random sampling frame that would give every document in the collection an equal chance of selection. The decision was made to take 0.001% of all the documents available, which totaled a little over 300 documents. Then specific month/year combinations were randomly selected and queried within the Tobacco Documents database to find out how many documents were available for selection. After the random selections were finished, all of the documents in the core corpus were classified using both linguistic and extralinguistic categories, including:

1. Public Health: Significant for Public Health or not significant for Public Health.
2. Audience: Industry-Internal Audience or Industry-External Audience was established to be exclusive of each other. Documents were classified as internal if they were addressed to persons or groups within or hired by the company from which the document originated, or if they were correspondence between tobacco companies. This was eventually extended to include vendors at all levels of the tobacco industry and all for-profit and for-hire organizations involved in the

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

research, growing, processing, distribution, and sale of tobacco products. Otherwise documents were classified as EX.

3. Addressee: Named or Unnamed.
4. Text Types. [24]

These criteria were used as the basis for making sure the contents of the corpus matched the intended use of the model. For example, all of the documents that were not designated as being significant for Public Health, being addressed to an industry-internal audience, or possessed a named addressee were rejected from becoming a part of the final quota sample. After creating the core sample for the TDC, the researchers used the distributions they observed to develop a protocol for sampling documents that fit their criteria to come a part of the quota sample. What they discovered was that their sampling process yielded proportions for document rejection were nearly the same for the final reference corpus as the initial pilot sample—although they were unable to verify these findings statistically to confirm reproducibility of the method.

As of this point, it is unknown if the principled method for subsampling Big Language Data outlined in "Looking for the Smoking Gun" is reproducible for a different monitor corpus. If it is reproducible, this particular method could be of critical importance to modeling Big Language Data, as it provides a means for actually measuring target populations of interest that are complex systems. In this paper, the role of principled sampling for creating corpora from Big Language Data resources addresses two specific aims:

1. How to operationalize the corpus creation model developed for the TIDs for a different, but similar, data set; and
2. Test whether the principled sampling method pioneered by the Tobacco Documents corpus is reproducible and if it does in fact provide maximal representativeness of a well-defined population of interest.

2.2. Applying Principled Sampling to Nuclear Power Discourse

Domain-specific language corpora are designed to represent language that serves a specific function, like the language of a particular industry. Most of these corpora are corporate in nature. While the study outlined in this article is based on the creation of a domain-specific corpus of regulated nuclear industry discourse, there is a more substantial, documented need for additional knowledge of sub-technical vocabulary for engineering disciplines for multiple contexts or extralinguistic points of scale [25]. The regulated nuclear power industry is, due to its complex regulatory history of efforts to increase public transparency and intra-industry learning after the Three Mile Island incident in 1979, an informative and novel case study for examining principled sampling techniques applied Big Language Data corpora.

The regulation of the nuclear industry began as a reaction to the use of atomic bombs on the Japanese cities of Hiroshima and Nagasaki in August of 1945. The United States Congress established the Atomic Energy Commission (AEC) by passing the Atomic Energy Act of 1946 in order to maintain control over atomic technologies and to investigate its military applications, and not necessarily to develop it for civilian purposes [26]. Following World War II, the primary focus of those individuals involved in nuclear development was directed toward military development. In the early part of 1953, the U.S. Navy began testing nuclear reactors to power their submarine fleet. After the Atomic Energy Commission observed the success of these reactors in autumn of the same year, it announced the intention to build a power plant. As a result, the first commercial nuclear reactor in the U.S. became operable in Shippingport, Pennsylvania, in 1957 [27]. Many more reactors would be built rather quickly in the years that followed.

The Atomic Energy Commission continued to regulate both the commercial use of atomic materials and the development of new technologies using those materials until Congress passed the Energy Reorganization Act of 1974, which divided the AEC into two agencies: the U.S. Energy Research and Development Administration and the U.S. Nuclear Regulatory Commission:

> The U.S. Nuclear Regulatory Commission (NRC) was created as an independent agency by Congress in 1974 to enable the nation to safely use radioactive materials for beneficial civilian purposes while ensuring that people and the environment are protected. The NRC regulates commercial nuclear power plants and other uses of nuclear materials, such as in nuclear medicine, through licensing, inspection and enforcement of its requirements. [28]

Thus, the NRC came into being in January 1975 to facilitate, and speed up, the licensing of nuclear plants, as well as to develop better regulatory practices for this industry. The issue of reactor safety is thought to be the central one for the NRC in its early years. One event, in particular, brought the safety of nuclear power plants, as well as the NRC, to the attention of the public, and that was an event known in the industry as the Brown's Ferry Fire:

> The first event was a major fire at the Tennessee Valley Authority's Browns Ferry Nuclear Plant near Decatur, AL, in March 1975. In the process of looking for air leaks in an area containing trays of electrical cables that operated the plant's control room and safety systems, a technician set off a fire. He used a lighted candle to conduct the search, and the open flame ignited the insulation around the cables. The fire raged for over 7 hours and nearly disabled the safety equipment of one of the two affected units. [26]

Only four years after this incident, another accident occurred at an American nuclear power generating station:

> On March 28, 1979, an accident at the Three Mile Island Nuclear Station (TMI), Unit 2, near Harrisburg, PA, made the issue [the risks of nuclear power] starkly and alarmingly real. As a result of a series of mechanical failures and human errors, the accident (researchers later determined) uncovered the reactor's core and melted about half of it.... By the time that experts realized that the plant had undergone a loss-of-cool- ant accident and flooded the core, the reactor had suffered irreparable damage. [26]

The rapid succession of the Brown's Ferry Fire and Three Mile Island affected the credibility of the nuclear power industry and the NRC, to put it lightly. However, in the years to come, this agency would develop safety requirements and regulatory practices that would help to reduce the risk and likelihood of future accidents.

As part of the Freedom of Information Act of 1966, the American public has a "right to know" about government records and documents [29]. Since September 11, 2001, the NRC provides to the public all documents about nuclear reactors here in the United States that are not found to contain "sensitive information." The NRC defines sensitive information as being data that has been found to be potentially useful to terrorists, proprietary knowledge for licensees, or "information deemed sensitive because it relates to physical protection or material control and accounting" [30]. All documents that do not possess these characteristics are made available through the NRC's Agency Documents Access and Management System (ADAMS) database (https://adams.nrc.gov/wba/).

ADAMS is composed of two secondary collections. First, there is the Publicly Available Records System (PARS) Library that "contains more than 730,0000 full-text documents that the NRC has released since November 1999, and several hundred new documents are added each day" [31] to a web-based archive. The second library is known as the Public Legacy Library and contains over 2 million bibliographic citations for documents earlier than those found in PARS.

In order to create a reference corpus of regulated nuclear power language from the ADAMS database, which is essentially a large monitor corpus, the Tobacco Documents Corpus methodology for assembling a pilot corpus was followed [32]. First, a different month for each of the 12 full years available as part of the ADAMS-PARS archive was randomly selected: 2000 through 2011 (Table 1).

Table 1. ADAMS Random Month Selection.

Year	Random Month Selection
2000	November
2001	January
2002	July
2003	September
2004	June
2005	February
2006	May
2007	August
2008	March
2009	October
2010	December
2011	April

The database was queried for each NRC licensee by using their docket numbers. Docket numbers are unique identification codes assigned to each licensee. All documents written by the licensee, written to the licensee, or sent to the licensee as informed communication for regulatory action or rulemaking are assigned to the licensee's docket. Primarily, the docket is considered a living record of communication for the licensee. As such, this identification number proves to be the ideal way for querying the available documents for each nuclear reactor regulated by the NRC. After the queries were finished, it was observed that this database performed similarly to that of the TIDs: the documents varied greatly in count and length for each month/year and each license (Table 2).

Table 2. ADAMS Document Availability by License Excerpt.

Year	Arkansas Nuclear 1	Beaver Valley 1	Braidwood 2	Browns Ferry 3	Byron 1
2000	20	9	25	12	17
2001	21	13	28	15	28
2002	21	11	25	22	7
2003	15	22	12	19	25
2004	21	15	9	22	10
2005	19	41	11	18	10
2006	16	15	40	29	22
2007	15	150	13	24	15
2008	11	32	25	16	19
2009	7	16	19	18	12
2010	6	3	12	12	14
2011	17	11	17	18	26

It was also determined that a sampling of 0.001 of all the documents available based on the initial querying would be taken, which totaled 30 documents per docket. These 30 documents were randomly selected across all 12 years based on the number of documents available within each year. An example of the sampling distribution for Indian Point 2, one of the 104 licensed nuclear reactors in the United States of America, can be found in Table 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2018 doi:10.20944/preprints201812.0016.v1

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

Table 3. Production-Based Document Sample for Indian Point 2.

Year	Random Month	Available	Sampled
2000	November	89	4
2001	January	95	4
2002	July	37	2
2003	September	31	1
2004	June	29	1
2005	February	10	1
2006	May	45	2
2007	August	121	5
2008	March	83	4
2009	October	42	2
2010	December	45	2
2011	April	50	2

After establishing the number of documents to be taken from each year for each licensee, random sets of integers were generated to represent each result from the query that would be selected as part of the pilot corpus. For example, the random selections for April 2011, for Indian Point 2 were entries 28 and 39. After the random selections were chosen, the appropriate documents were downloaded from ADAMS as .PDF files that had already been converted into a machine-readable format using optical character recognition (OCR) software by NRC librarians.

One of the advantages of leveraging the NRC ADAMS database as a Big Language Dataset for subsampling is that there are extensive metadata about each document. (Figure 1). Within the ADAMs database, users can select exactly which metadata fields are needed for classifying documents, while also exporting the chosen fields and entries to .CSV files. Metadata fields such as Document Type, Author Affiliation, Addressee Affiliation, and even the originating Docket Number of the documents are provided for this database. A .CSV file was exported for all Pilot selections to expedite document classification.

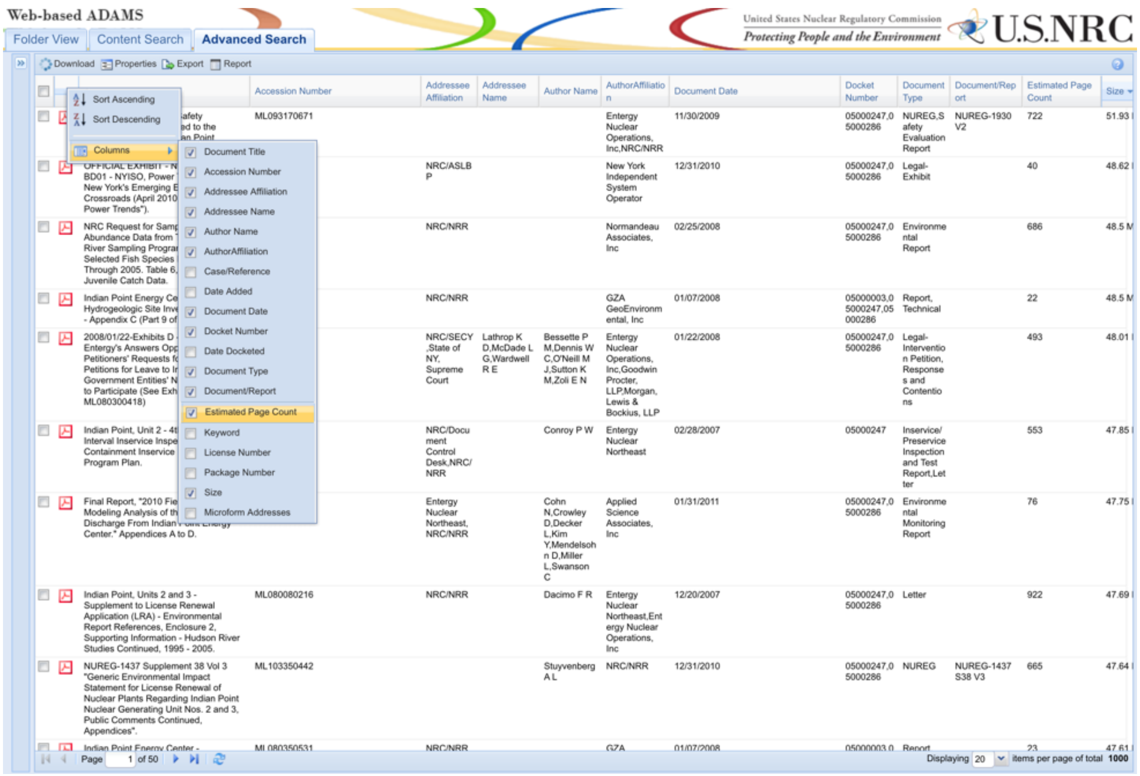

Figure 1. Adams report selection.

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

Comparing all of the metadata provided for the randomly selected documents in the pilot to their requisite .PDF files, the resulting samples were classified according to the following guidelines adapted from those used to create the Tobacco Documents Corpus:

1. Nuclear Power Regulation: No communications involving the regulation of nuclear materials for medical or research uses were included in the pilot corpus, only documents related to the regulation of nuclear power.
2. Industry-Internal Author/Audience or Industry-External Author/Audience: Documents are classified as Audience Industry-Internal if they are addressed to persons or groups within or hired by the licensees or the NRC, or if the document is correspondence between individuals at the NRC or individual licensees. Furthermore, vendors at all levels of the nuclear industry and all consultants (legal, environmental, etc.) and contractors (engineering firms) involved in the production, management, regulation, or business of nuclear power are to be considered internal as well. Otherwise documents are classified as external to the nuclear power industry.
3. Document Types: All documents are assigned document type designations by the NRC librarians. These designations can be found on the Custom Legacy report.
4. Docket Designation: If the docket number assigned to the document is the same as the licensee, it was classified as "Own." The designation "Other-Same Site" was used if the docket number was that of a licensed nuclear reactor on the same site. "Other-Same Corporation," designated the situations where the originating docket number assigned to the document represents a licensee owned by the same corporation as the docket number being searched for each document. Finally, the designation "Other-No Affiliation," was used to indicate documents assigned to a licensee's docket that originated from a licensee not possessing any of the aforementioned qualities.
5. Language-Based: All of the documents are marked as being language-based or not in order to identify documents that are image-based like drawings and photographs.
6. Length: Texts shorter than 50 words of continuous discourse were marked so that they can be excluded from the corpus. Likewise, documents longer than 3,000 words are denoted in the metadata so that they can be sampled (1,000 words from the beginning, 1,000 words from the middle, and 1,000 words from the end) to avoid bias.

Once all of the classifications for the pilot corpus were made, selection compliance with the sampling framework was performed in order to identify characteristics of the documents sampled from the population of those available to the public on the ADAMS Database.

3. Results

One of the first observations made through document classification process for the Pilot was that although the sample only allowed for unique document selections of the results from each docket number's database query, duplicate documents (documents being assigned identical accession numbers by the NRC) were sampled because a single document may be assigned to multiple dockets by the NRC. By reconciling the metadata provided by the database for each document randomly selected to be part of the corpus with the sampling framework, the exact dockets assigned to a specific document were able to be identified. For the purpose of the reference corpus, this particular occurrence distorted the sampling of the pilot at the docket level due to over-representation of certain documents. However, the inter-docket relationships of documents in this corpus needed to be preserved as it contributes to potential shared language of multiple licensees, albeit utilizing sampling with replacement statistics. As a result of eliminating all of the duplicate documents from the Pilot, the 3,120 documents downloaded from the ADAMS were reduced to 2,775 unique samples.

Another characteristic documented by the NRC librarians within the ADAMS database is document type. Concerning the types of documents that are part of the Pilot sample, an interesting pattern emerges the aggregate frequencies are plotted. As is seen in Figure 2, there is a very distinct, and steep, asymptotic hyperbolic curve, or A-curve.

Peer-reviewed version available at *Data* **2019**, *4*, 48; doi:10.3390/data4020048

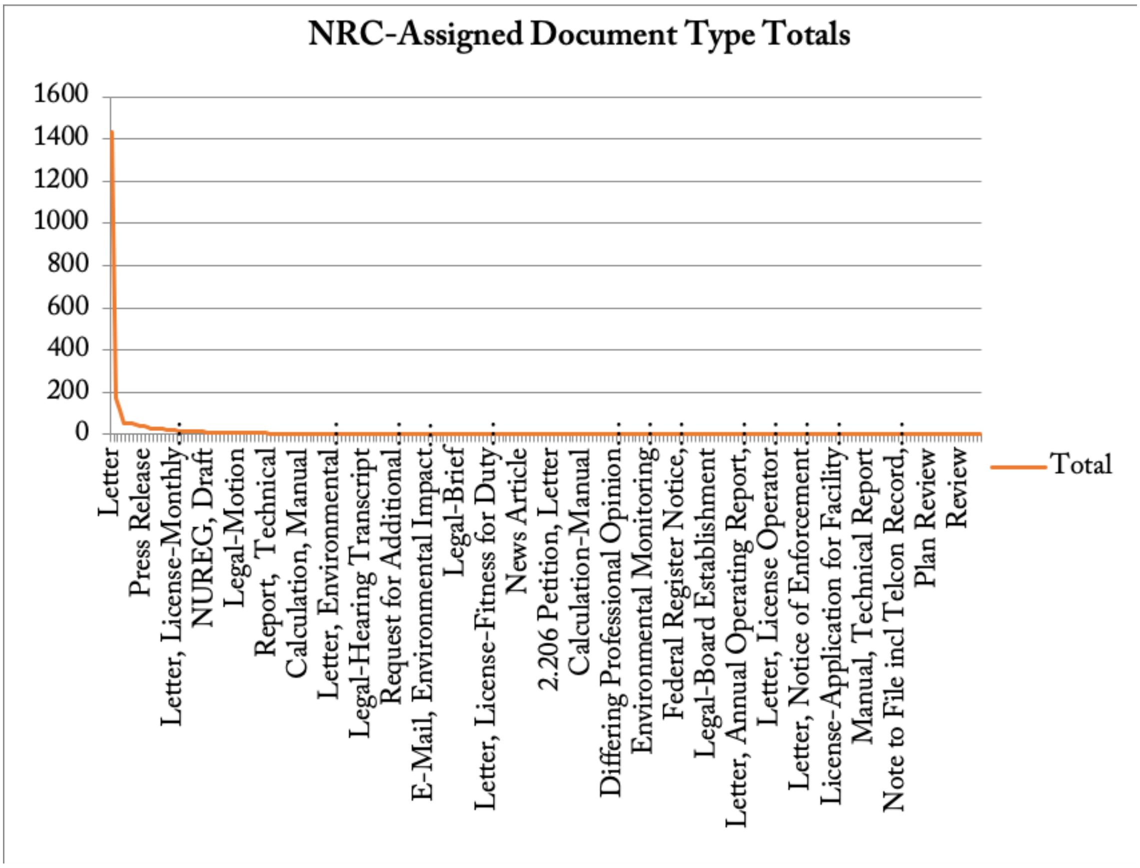

Figure 2. Pilot Document Totals Before Splitting Multiples.

In the case of the data in Figure 2, word frequencies are not plotted against their ranks, but rather document types. For the Pilot, it can be seen that the NRC has denoted a majority of the documents as being letters, 1,125 in fact. However, when looking at these documents, many of them appeared to be rather long. So, each document was visually verified and coded for whether or not they had a unique attachment: 44.45% of them did. Because of this observation, although the NRC librarians have designated a particular file as being a specific document type when it comes to letters especially, the potential exists for multiple document types to be present. After splitting these multiple documents, the result was 4,773 individual .PDF files in the sampling. Once all of the files possessing multiple documents were split apart, thereby changing the scale of document types in the Pilot, there still appears to be an A-curve with regard to the relative frequencies of the document types (Figure 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2018 doi:10.20944/preprints201812.0016.v1

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

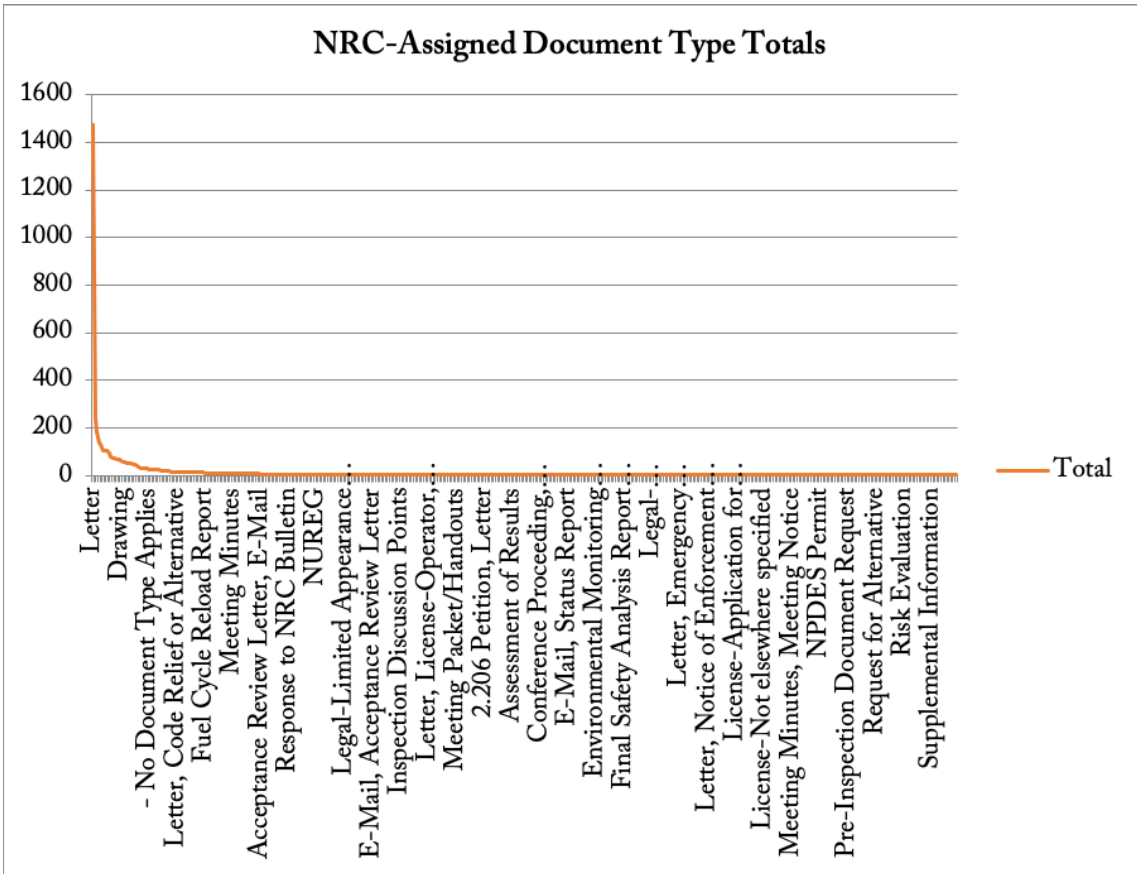

Figure 3. Pilot Document Totals After Splitting Multiples.

Letters were still the most common document after the scale changed, but the frequencies of other documents like Safety Evaluations increased drastically (from 2 to 104). Although the number of document types in the Pilot changed, as well as their relative distribution, the A-curve is still present. This particular behavior is called scaling: the A-curve is present at different aspects, or levels of scale, in the corpus. Scalability of data through A-curve distributions has also been documented extensively in speech data across different linguistic variables, time, and even geographic locations (W. A. Kretzschmar 2009). The frequency of document types is, in fact, scalable for this particular population of documents. This characteristic is an essential quality of language in use that should also be documented in the lexical frequencies of the ADAMS documents concerning proximity.

In order to learn more about the language of the nuclear industry, not only do the documents in the corpus need to be about nuclear power, but also the authors need to be classified as internal. Of the 4,773 documents from the ADAMS-PARS database, 97.76% of them were authored by internal sources. Thus, 4,666 documents were kept as part of the reference corpus while 107 documents were not (externally affiliated authors wrote 105 of these documents, and the affiliation of two documents could not be determined). Concerning the internal/external status of the sampled documents' audience affiliations, since the function of the NRC is to ensure "that people and the environment are protected, (NRC 2016)" both internally and externally-directed documents are maintained as part of the corpus.

Of the 4,666 documents remaining in the Pilot, only 2.27% (or 106 of them) were not language-based documents, such as drawings and photographs (Figure 4).

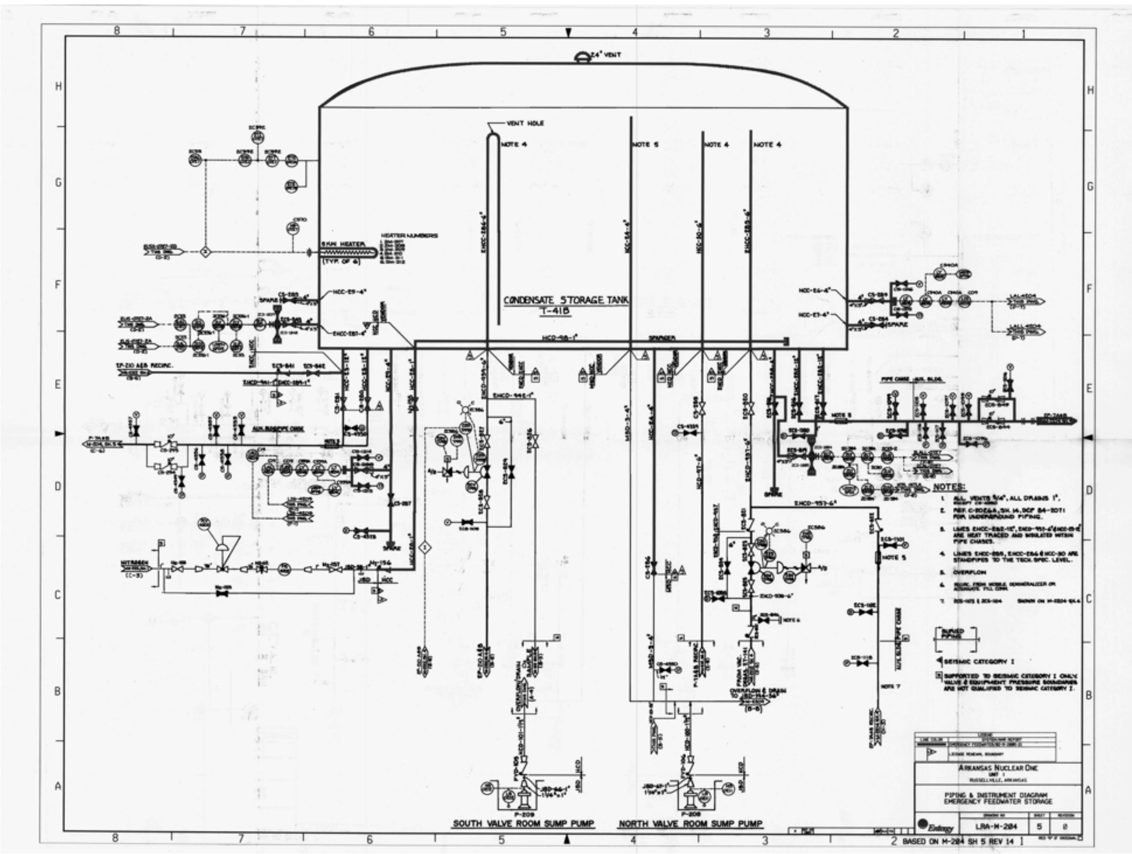

Figure 4. Arkansas Nuclear One Condensate Storage Tank Drawing.

They were not kept as part of the reference corpus. For the 4,560 documents now remaining in the pilot, the average page length was 32.3 pages with a standard deviation of 79.79. The length of the documents available from the NRC database is highly variable with documents ranging from one page to 2,996 pages. However, just because a document has numerous pages does not necessarily mean that it contains a great many words. When looking at the sampled documents, 78.79% of them (3,806) contained 50 words or more of continuous discourse. As a result, 967 documents could not be used because they were too short. After taking out all of the documents from the pilot sample that were not authored by groups internal to the nuclear power industry, were not language based, and had less than 50 words of continuous discourse, we were left with 3,593 documents. In other words, the Pilot had a rejection rate of 24.72

In order to see if this random selection methodology was fruitful and yielded reproducible and consistent results, three additional iterations of the sampling protocol were performed to look for consistency in the proportions of document rejection to create a sizable reference corpus from the ADAMS database.

4. Discussion

One of the essential qualities of a sampling methodology is that it be reproducible. For this reason, three additional rounds of sampling were performed with the NRC ADAMS database using the previously described protocols. One way to evaluate the reliability of this sampling method is to evaluate the statistical similarities, or instead evaluate if there are any differences statistically in the rates of rejection for documents in the second, third, and fourth iterations of sampling with respect to the Pilot for all of the classification criterion. Although a quota-derived sampling protocol based on the documents available in the ADAMS database was used, it was necessary to verify whether or not the ratios of documents rejected due to the qualities of each document were consistent across all of the iterations in comparison to the Pilot.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2018 doi:10.20944/preprints201812.0016.v1

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

In order to evaluate the sampling procedures, a two-proportion z-test at a 99% confidence level was performed at each stage where documents were rejected. As was done with the Pilot, all of the files that were duplicates for their unique Accession identification numbers for each iteration were eliminated. There was no statistically significant difference between the rejection ratios of all three iterations in comparison to the pilot (Table 4).

Table 4. Duplicate Accession ID Rejection Ratios.

. Iterations	Duplicate Documents	Total Documents	Rejection Ratio
Pilot	345	3120	11.06%
Iteration 2	355	3120	11.38%
Iteration 3	368	3120	11.79%
Iteration 4	371	3120	11.89%

After making sure all of the documents within each iteration were represented only once, all files were verified to be composed of only one document. The resulting proportions of documents also had no statistical difference from the pilot at a 99% confidence level (Table 5).

Table 5. Ratio of Original Number to Number After Splitting Multiples.

. Iterations	Original Number of Documents	Number of Documents After Split	Ratio
Pilot	2775	4773	58.14%
Iteration 2	2765	4625	59.78%
Iteration 3	2752	4618	59.59%
Iteration 4	2749	4581	60%

There was still no statistically significant difference between the rejection ratios of all three iterations in comparison to the Pilot after eliminating all duplicates, splitting all files possessing multiple documents, and eliminating all of the externally-authored documents (Table 6).

Table 6. Externally-Authored Document Rejection Ratios.

. Iterations	Externally-Authored Documents	Total Documents	Rejection Ratio
Pilot	107	4773	2.24%
Iteration 2	111	4625	2.4%
Iteration 3	90	4618	1.95%
Iteration 4	106	4581	2.31%

After all of the externally-authored documents were removed from the sampling for each iteration, all of the remaining documents classified as not being language-based were also filtered out. Again, the proportion of internally-authored documents that were not language-based was consistent across all three additional iterations in comparison to the Pilot at a 99% confidence level (Table 7).

Table 7. Non-Language-Based Document Rejection Ratios.

. Iterations	Non-Language-Based Documents	Total Documents (Internally-Authored)	Rejection Ratio
Pilot	106	4666	2.27%
Iteration 2	113	4514	2.5%
Iteration 3	104	4528	2.3%
Iteration 4	103	4475	2.3%

The final step for all three of the additional iterations was to identify all of the documents having at least 50 words of continuous discourse. Using the database metadata, the number of documents that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2018 doi:10.20944/preprints201812.0016.v1

Peer-reviewed version available at Data 2019, 4, 48; doi:10.3390/data4020048

were internally-authored and language-based, but too short for inclusion according to the classification criteria, were verified. With a 99% confidence level, not only was it verified that these proportions also did not have a statistically significant difference for this final classification (Table 8), but also concerning the total rate of rejection for iterations two through four in comparison to the pilot sample (Table 9).

Table 8. Document Length Rejection Ratios.

. Iterations	Documents Having Fewer Than 50 Words	Internally-Authored & Language-Based	Rejection Ratio
Pilot	967	4560	21.21%
Iteration 2	886	4401	20.13%
Iteration 3	865	4424	19.55%
Iteration 4	831	4372	19.01%

Table 9. Total Rejection Ratios for All Iterations.

. Iterations	All Documents Rejected	Total Documents	Rejection Ratio
Pilot	1180	4773	24.72%
Iteration 2	1110	4625	24%
Iteration 3	1059	4618	22.93%
Iteration 4	1040	4581	22.70%

This analysis provides an additional level of confidence that the sampling procedure outlined in "Looking for the Smoking Gun," is reliable across multiple iterations, reproducible, and yields a consistent and representative model of the population of interest defined by the sampling framework.

5. Conclusion

The findings of this study, while demonstrating that the Tobacco Documents Corpus principled sampling method is a valid one, corroborate recent studies claiming that even Big Language Data corpora should not be considered as a black box as any subsampling of extralinguistic factors from an existing reference corpus could ignore within-group variation [33]. Thus, there is a distinct opportunity for future research around designing corpora from Big Language Data that exhibit characteristics of complex systems. Extralinguistic factors and linguistic characteristics of documents sampled in the creation of corpora have the potential to be highly interconnected and should be further investigated. Blending a principled sampling framework with demographic sampling in the next iteration of corpus sampling through human-centered design would address this opportunity by facilitating the use of techniques that shift the focus to the people involved in the creation of linguistic data, rather than language as the sole artifact of interest for analysis.

Acknowledgments: I would like to acknowledge the amazing work performed by the Metadata Librarians of the Nuclear Regulatory Commission, without whose work on the ADAMS database this project would not be possible.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TID	Tobacco Industry Document
TDC	Tobacco Documents Corpus
AEC	Atomic Energy Commission
NRC	Nuclear Regulatory Commission

References

1. Lohr, S. The age of big data. *The New York Times* **2012**, 1–5. doi:10.1126/science.1243089.
2. Johansson, S. Times change, and so do corpora. In *English Corpus Linguistics.* Routledge,**2014**, 305.
3. Johansson, S., and Stenström, A. *English Computer Corpora: Selected Papers and Research Guide.* **1991**, 3. Walter de Gruyter.
4. Baker, P. *Using Corpora in Discourse Analysis.* **2006**, A&C Black.
15. Crawford, K. Six provocations for big data. **2011**, 1–17. doi:10.2139/ssrn.1926431
6. Manovich, L. Trending: The promises and the challenges of big social data. In *Debates in the Digital Humanities.* Minneapolis, MN: University of Minnesota Press. **2011**, 460–75.
7. Davies, M. *TIME Magazine Corpus (100 Million Words, 1920s-2000s).* (Accessed November 2018) http://corpus.byu.edu/time
8. Kilgarriff, A., and Grefenstette, G. Introduction to the special issue on the web as corpus. *Computational Linguistics.* **2003**, 29(3). 333–47. doi:10.1162/089120103322711569
9. Introna, L.D., and Nissenbaum, H. Shaping the web: Why the politics of search engines matters. *The Information Society.* **2000,** 16(3). 169-185.
10. Meyer, C.F., and Nelson, G. Data Collection. In *The Handbook of English Linguistics.* **2006,** 36. Wiley-Blackwell. 93.
11. Kennedy, G. *An Introduction to Corpus Linguistics.* **1998.**London: Longman.
12. Blackwell, M., and Sen, M. Large datasets and uou: A field guide. *Manuskript 14627.* **2012,** 1–8. (Accessed in November 2018) http://www.mattblackwell.org/files/papers/bigdata.pdf.
13. Sinclair, J. Corpus and text–Basic principles. In *Developing linguistic corpora: A guide to good practice.* Ed. by Martin Wynne. (Accessed in November 2018) http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm.
14. Kretzschmar, W.A. *Language and complex systems.* **2015**, Cambridge University Press.
15. Crawford, K. 2011. Six provocations for big data. **2011,** 1–17. doi:10.2139/ssrn.1926431.
16. Kretzschmar, W.A. Language variation and complex systems. *American speech.* **2010,** 85(3). 263-286.
17. Burkette, A. The lion, the witch, and the armoire: Lexical variation in case furniture terms. *American speech.* **2009,** 84(3). 315-339.
18. Lohr, S.*Sampling: Design and Analysis.* Cengage Learning.**2009.**
19. Kretzschmar, W.A., Meyer, C.F., and Ingegneri, D. Uses of inferential statistics in corpus studies. *Language and computers.* **1997,** 20. 167–78.
20. Meyer, C.F. *English corpus linguistics: An introduction.* Cambridge University Press. **2002.**
21. Biber, D. Representativeness in corpus design. *Literary and linguistic computing.* **1993,** 8(4). 243–57. doi:10.1093/llc/8.4.243.
22. Crowdy, S. Spoken corpus design. *Literary and linguistic computing.* **1993,** 8 (4). 259–65.
23. Kretzschmar, W.A., Darwin, C., Brown, C., Rubin, D. and Biber, D. Looking for the smoking gun: Principled sampling in creating the Tobacco Industry Documents Corpus. *Journal of english linguistics.* **2004,** 32(1). 31–47. doi:10.1177/0075424204263024.
24. Kretzschmar, W.A. Sampling plan for creation of corpora for the Tobacco Documents Grant. **2001.**
25. Mudraya, O. Engineering English: A lexical frequency instructional model. *English for specific purposes.* **2006,** 25(2). 235–56. doi:10.1016/j.esp.2005.05.002
26. Walker, S.J., and Wellock, T.R. A short history of nuclear regulation, 1946-2009. *U.S. Nuclear Regulatory Commission.* 2010.
27. Bodansky, D. *Nuclear energy: Principles, practices, and prospects.* New York: Springer Publishers.**2004.**
28. The United States Nuclear Regulatory Commission. About NRC. *The United States Nuclear Regulatory Commission.* **2018,** (Accessed in November 2018) https://www.nrc.gov/about-nrc.html
29. Henry, C.L. *Freedom of Information Act.* Nova Publishers. **2003**
30. The United States Nuclear Regulatory Commission. Withholding of sensitive information for nuclear power reactors. *The United States Nuclear Regulatory Commission.* **2018,** (Accessed in November 2018) http://www.nrc.gov/reading-rm/sensitive-info/reactors.html.
31. The United States Nuclear Regulatory Commission. ADAMS Public Documents. *The United States Nuclear Regulatory Commission.* **2018,** (Access in November 2018) http://www.nrc.gov/reading-rm/adams.html.

32. Hettel, J. 2013. *Harnessing the power of context: A corpus-based analysis of variation in the language of the regulated nuclear industry.* **2013,** Athens, GA: University of Georgia dissertation. (Accessed in November 2018) https://getd.libs.uga.edu/pdfs/hettel_jacqueline_m_201305_phd.pd.
33. Březina, V., and Meyerhoff, M. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International journal of corpus linguistics.* **2014,** 19(1). 1–28. doi:10.1075/ijcl.19.1.01bre.