data

MDPI

Article

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

Jacqueline Hettel Tidwell

Department of English, Franklin College of Arts and Sciences, University of Georgia, Athens, GA 30602, USA; jacqueline.tidwell@uga.edu

Received: 29 November 2018; Accepted: 30 March 2019; Published: 2 April 2019

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. After performing multiple sampling iterations, the approach pioneered by the Tobacco Documents Corpus creators was deemed to be reproducible and valid using a two-proportion z-test at a 99% confidence interval at each stage of the evaluation process–leading to a final mean rejection ratio of 23.5875 and variance of 0.891 for the documents sampled and evaluated for inclusion into the final text-based model. 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.

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 [5,6]. Big Language Data corpora are not merely 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.

The development, assessment, and dissemination of principled subsampling methods for designing and constructing Big Language Data corpora from existing sources is a topic that is deserving of critical inquiry given such scrutiny underlying unsampled Big Language datasets for understanding social issues. The process of critically analyzing the creation of these models from a decision-making perspective is important for the diffusion and adoption of sampling methods more broadly in other disciplines, within which thought leaders are calling for more transparent and critical treatments of research methods and research design of sociotechnical issues like energy [12]. Therefore, the purpose of this study is to provide an alternate perspective of text-based corpus creation as an interpretive act, inflected by the theoretical position and perspective of its designer, and defined by the nature and extralinguistic factors that precipitated the documents' creation. All of the effort and planning that goes into the design of corpora for understanding language as it is really used—especially when we are dealing with Big Language data—is an aspect of semantic research that is often overlooked. Thus in this paper, a novel approach for demographic sampling of large-scale databases to create reference corpora that was pioneered in the wake of a public health crisis in the United States is presented and described in detail. The application of this method to a separate database, the U.S. Nuclear Regulatory Commission's ADAMS database, gave the opportunity to test the reproducibility and validity of this approach over multiple iterations. Finally, this article will consider and test this method of Big Language Data downsampling of digitally accessible language data for its reproducibility and automation in future research.

2. Background

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 [13]. 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" [14]. 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 [15]. 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 [5]. 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 [15–17]. 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 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. Corpus linguistics is one research methodology that, while an exercise in modeling, allows the use of "real-life language" sampled from the world in which it is used. McEnery and Wilson [18] define corpus linguistics as "the study of language based on examples of real-life language use." However, other scholars in this field like John Sinclair [14], argue that corpus linguistics is instead a 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." Why is it important then to create a model that is composed of samples of real language, texts if you will?

Language, social action, and knowledge all coexist together. In fact, the way in which words are used "can reveal relations between language and culture: not only relations between language and the world, but also between language and speakers with their beliefs, expectations and evaluations" [19]. Whether or not we are conscious of it, we have these expectations for the language we use everyday. For example, when we read a newspaper article, we expect the words and phrases it uses to be quite different from those used in a technical manual for putting together a child's toy. That is, there are a diverse set of linguistic expectations for each text-type to which we have been exposed, and as such the creation of a computationally tractable model of language necessitates a formalized plan for how it is created.

Language behavior of the texts that we might select for our model or corpus, will vary by text type. Text types are considered to be situationally recognizable speech acts that range from large to small in scale: e.g., all written language versus a letter or even a job application letter [20]. Our expectations vary from text-type to text-type, and we easily recognize that the language used in different text-types, and thus the documents that are examples of them, are different. The ways in which the interaction of extralinguistic contexts with linguistic elements correlate with differences in the way meaning is created within different text-types is a reality that corpus builders must contend with when creating corpora for Big Language data analysis so as to maintain its computational tractability.

Corpus linguistics is a form of modeling through computing that enables us to specify computationally "how we know what we know", and we need to marry an explicit model with intuition because we can know more than we can tell [21]. That being said, corpora cannot provide negative evidence about a language (i.e., what is correct or possible, or what is incorrect or not possible). A corpus cannot tell you why certain patterns occur in language; as we have already discussed, this is where intuition comes into the picture. What a corpus can do is tell you what happens within it, and with statistics it can help you understand the propensity for those things to happen. Finally, a corpus cannot provide all of the possibilities in language at one time because it must be principled, it must be planned, and it must be systematically constructed [22]. This importance on constructing a corpus that is representative of the type of language desired for analysis–even Big Language Data Corpora–is based on the need for manipulation of a model that yields accurate interpretations of what we are modeling and not something else entirely.

2.1. Related Works

One tool for defining specific subsets of language-based data for inclusion in a corpus is through the use of a sampling policy or 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 [23]. Generally, there are three primary considerations that must be addressed in establishing a sampling policy or framework: the orientation of the language or variety to be sampled; the criteria upon which the samples will be chosen; and the nature and dimensions of the samples [14]. 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 [22,24]. Employing such an approach affords a linguist the confidence that the corpus is representative of the complex system they are attempting to model.

2.1.1. Content and Orientation

The notion of the orientation of a corpus is a consideration that is often taken for granted by many corpus builders. For a researcher to make claims about the authenticity of results coming from the analysis of a particular corpus or sampling of documents they must make sure that "only those components of corpora which have been designed to be independently contrastive should be contrasted" [14]. In other words, the type of corpus you choose to create dictates how you sample texts for inclusion in the corpus.

The type of corpus to be created is dependent on the kind of analysis, or manipulation to be performed [25]. More specifically, it is dependent on what you want to study (content) and the context of that content (orientation). Table 1 lists a few types of corpora as defined by their content and orientation.

Table 1. Corpus type by content and orientation.

Content	Orientation
General	Represents a language or language variety as a whole.
Balanced/Representative	Texts are selected using pre-defined sampling proportions.
Historical	Represents an earlier stage(s) of a language.
Monitor	New texts are added to the corpus continuously to "monitor" language change.
Regional	Represents one regional variety of a language.
Parallel	The same texts are selected in two or more languages or language varieties.
Learner	Represents the language produced by learners of a particular language.
Comparative	Similar texts are selected in two or more languages or varieties.
Reference	Represents a very specific type of language.
Diachronic	Texts from consecutive time periods are included for comparison purposes.

Once the content has been determined for this linguistic model, a corpus builder also needs to decide the context of interest for the inquiry. Depending on the contrastive elements of the inquiry, a corpus builder should orient the sampling of their corpus in different ways. In other words, are they trying to make observations to define a particular language or variety? If so, then they are creating a balanced/representative corpus and should use predefined sampling proportions. Is their real interest in monitoring how a particular language has changed over time? If so, then their corpus is a monitor one and should continuously add new texts over time, starting with the year corresponding with the beginning of her time period of interest. Each of these different types of corpora create a situation where researchers can make observations about a language by contrasting particular elements: i.e., time, extralinguistic parameters, text-type, etc.

2.1.2. Representativeness

Representativeness is an important aspect of corpus creation. There is no way that one can make accurate generalizations about a language or language variety if the corpus is not sampled validly from the language or language variety in question using a sampling policy that assists in the determination of which types of texts can be included in a corpus. These criteria all focus on the nature of a text: its mode, type, domain, language, location, and even date. He suggests that these criteria be "small in number, clearly separate from each other, and efficient as a group for delineating a corpus that is representative of the language or variety under examination" [14]. This issue of representativeness is defined within the sampling parameters established before the corpus is created and has the potential to significantly impact the observations made from it [26–29].

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 [28,30]. 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 [31]. For demographic sampling, data to comprise 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.

The decisions made initially regarding which texts are included or excluded are of great importance to the creation of a corpus, as they directly influence its representativeness and the models generated from it. This issue of representativeness has become of increasing interest in light of the popularity for using the Google Books Corpus and other born digital resources in language research [32]. Because the Google Books Corpus is essentially a library of indexed texts, and was collected using convenience and snowball methods, researchers have been seduced into interpreting false analyses of the language contained within it. Recent research calls for the need to better understand the composition and dynamics–the backstory, if you will–of the corpus before using it to generate broad conclusions about the evolution of languages and cultures [33].

2.1.3. Dimension and Scale

The third and final consideration for establishing a sampling policy for a corpus, as outlined by Sinclair, relates to the nature and dimensions of the samples. In other words, how many texts are needed in the corpus in order to create valid estimates: how big does your model need to be? This question is connected directly to the issue of what kind of manipulations you want to do with your model of language in use. The focus of size in corpus construction, even the creation of Big Language Data corpora, is not necessarily to create the largest corpus possible, but rather to create a model that will facilitate the analysis and inquiry for the purposes needed for the investigation in question.

There are two general types of population sampling that also influence the size of a corpus: probability and non-probability sampling. Probability sampling is defined by a researcher carefully pre-selecting the population she wants to study with her corpus by using statistical formulas and demographic information to ensure representativeness [22]. On the other hand, non-probability sampling is a process during which pre-selection is not employed. We can generally think of there being five general types of non-probability sampling for language-based research:

1. Haphazard, convenience, or accidental sampling, where a researcher only samples those individuals who are available.

2. Self-selection sampling, where participants choose whether or not they take part.
3. Snowball sampling, where participants are difficult to reach and thus must be recruited based off of existing networks of relationships.
4. Judgment, or purposive, or expert choice sampling, where a researcher decides ahead of time who is best qualified to be sampled (for example, only taking samples from native speakers of a language rather than non-native ones).
5. Quota sampling, where a researcher samples certain percentages of certain populations; for example, she could create a corpus whose samples reflect actual percentages of students of University Y with regard to gender (60% females, 40% males).

While it has been suggested that probability sampling is the most reliable type of sampling, because it leads to the least amount of bias, it is often logistically impossible for linguists to do this type of sampling because it often yields extremely large sample sizes, and by extension sizeable resources and funding [22]. As a result of these implications associated with probability sampling, it is common for corpus builders to use various combinations of non-probability sampling methods in creating corpora. Moreover, they are able to make valid estimates and observations of their intended populations through the use of documented sampling policies.

Once the population to be investigated with a corpus has been defined, as well as the sampling framework, there are generally two different ways in which the sampling can be performed: simple random sampling and stratified random sampling. Simple random sampling is where all of the sampling units within the sampling frame for a corpus are assigned a number and are then chosen at random using a random number generator or table. This type of sampling can be problematic, as "the chance of an item being chosen correlates positively with its frequency in the population, [and] simple random sampling may generate a sample that does not include relatively rare items in the population, even though they can be of interest to researchers" [30]. This approach is in contrast to a stratified random sample where the population of sampling units is divided into "relatively homogeneous groups (so-called strata) and samples of each stratum" are taken at random [30]. For example, if we were interested in sampling according to demographic factors, we could perform a stratified random sample where the population of sampling units are divided on the basis of the age, sex, and/or social class of the writers or speakers. According to Biber [26], a stratified sample is never less representative than a simple random sample; however, a simple random sample can be less representative than a stratified sample. These decisions yield great implications for the analysis and modeling of language-based data at any scale.

3. 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.

In 2004, W.A. Ketzschmar et al. [34] 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 [35]. 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 research, growing, processing, distribution, and sale of tobacco products. Otherwise documents were classified as EX.
3. Addressee: Named or Unnamed.
4. Text Types [35].

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.

4. Materials

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 [36]. 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 [37]. 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 [38]. 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. [39]

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. The rapid succession of the Brown's Ferry Fire in 1975, and Three Mile Island in 1979, affected the credibility of the nuclear power industry and the NRC [37]. 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 through multiple methods including data and information sciences.

As part of the Freedom of Information Act of 1966, the American public has a "right to know" about government records and documents [40]. Since 11 September 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" [41]. 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 7,300,000 full-text documents that the NRC has released since November 1999, and several hundred new documents are added each day" [42] 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 2010, the NRC introduced the Web-Based ADAMS system. This interface allows users the ability to search across both the PARS and Public Legacy libraries for publicly available documents such as regulatory guides, NUREG-series reports, inspection reports, Commission documents, correspondence, and other regulatory and technical documents written by NRC staff, contractors, and licensees.

Sampling Parameters

This system was assessed to be an ideal candidate for testing the TDC method, as it allows full-text searching and enables its users to view document images, download files, and export the relevant metadata. 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 [43]. 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 2).

Table 2. 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 3).

Table 3. 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

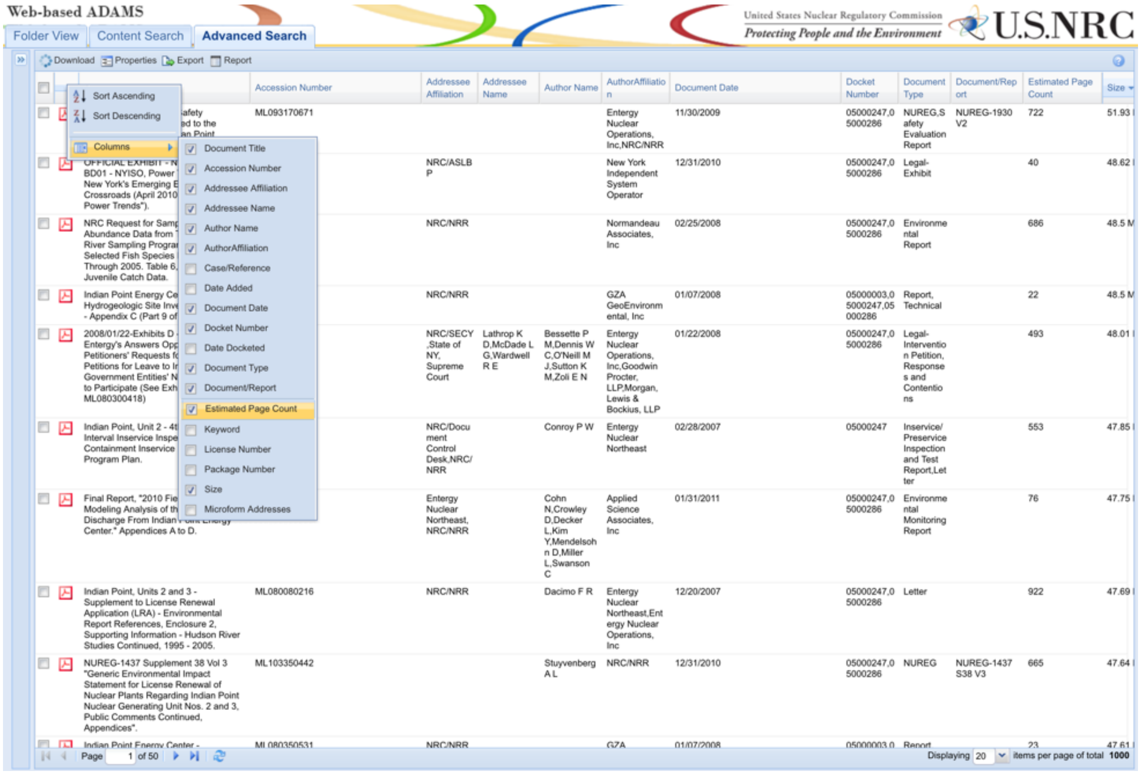
It was also determined that a sampling of 0.001 of all the documents available based on the initial querying would be taken, which totaled 30 documents per docket. These 30 documents were randomly selected across all 12 years based on the number of documents available within each year. An example of 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 4.

Table 4. 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).



	Accession Number	Addressee Affiliation	Addressee Name	Author Name	AuthorAffiliation	Document Date	Docket Number	Document Type	Document/Report	Estimated Page Count	Size
afety ed to the an Point	ML093170671				Entergy Nuclear Operations, Inc,NRC/NRR	11/30/2009	05000247,05000286	NUREG,Safety Evaluation Report	NUREG-1930 V2	722	51.93
OFFICIAL EXHIBIT - N BD01 - NYISO, Power New York's Emerging E Crossroads (April 2010 Power Trends").		NRC/ASLBP			New York Independent System Operator	12/31/2010	05000247,05000286	Legal-Exhibit		40	48.62
NRC Request for Samp Abundance Data from River Sampling Prograi Selected Fish Species Through 2005. Table 6, Juvenile Catch Data.		NRC/NRR			Normandeau Associates, Inc	02/25/2008	05000247,05000286	Environmental Report		686	48.5 M
Indian Point Energy Ce Hydrogeologic Site Inv - Appendix C (Part 9 of		NRC/NRR			GZA GeoEnvironmental, Inc	01/07/2008	05000003,05000247,05000286	Report, Technical		22	48.5 M
2008/01/22-Exhibits D Entergy's Answers Opp Petitioners' Requests fo Petitions for Leave to Ir Government Entities' N to Participate (See Exh ML080300418)		NRC/SECY,State of NY, Supreme Court	Lathrop K D,McDade L G,Wardwell R E	Bessette P M,Dennis W C,O'Neill M J,Sutton K M,Zoli E N	Entergy Nuclear Operations, Inc,Goodwin Procter, LLP,Morgan, Lewis & Bockius, LLP	01/22/2008	05000247,05000286	Legal-Intervention Petition, Responses and Contentions		493	48.01
Indian Point, Unit 2 - 4t Interval Inservice Inspe Containment Inservice Program Plan.		NRC/Document Control Desk,NRC/NRR		Conroy P W	Entergy Nuclear Northeast	02/28/2007	05000247	Inservice/Preservice Inspection and Test Report,Letter		553	47.85
Final Report, "2010 Fie Modeling Analysis of th Discharge From Indian Point Energy Center." Appendices A to D.		Entergy Nuclear Northeast, NRC/NRR		Cohn N,Crowley D,Decker L,Kim Y,Mendelsohn D,Miller L,Swanson C	Applied Science Associates, Inc	01/31/2011	05000247,05000286	Environmental Monitoring Report		76	47.75
Indian Point, Units 2 and 3 - Supplement to License Renewal Application (LRA) - Environmental Report References, Enclosure 2, Supporting Information - Hudson River Studies Continued, 1995 - 2005.	ML080080216	NRC/NRR		Dacimo F R	Entergy Nuclear Northeast,Entergy Nuclear Operations, Inc	12/20/2007	05000247,05000286	Letter		922	47.69
NUREG-1437 Supplement 38 Vol 3 "Generic Environmental Impact Statement for License Renewal of Nuclear Plants Regarding Indian Point Nuclear Generating Unit Nos. 2 and 3, Public Comments Continued, Appendices".	ML103350442			Stuyvenberg A L	NRC/NRR	12/31/2010	05000247,05000286	NUREG	NUREG-1437 S38 V3	665	47.64
Indian Point Energy Center -	ML080350531	NRC/NRR			GZA	01/07/2008	05000003,0	Report		23	47.61

Figure 1. ADAMS report selection.

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.

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 3000 words are denoted in the metadata so that they can be sampled (1000 words from the beginning, 1000 words from the middle, and 1000 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 by human raters in order to identify characteristics of the documents sampled from the population of those available to the public on the ADAMS Database. This six-step evaluation process, depicted as a process diagram in Figure 2, was performed on each document that was randomly selected for inclusion into the corpus. Before a document was deemed acceptable as a component of the final text-based model, it was evaluated regarding its fit to the research question's specifications regarding context, structure or text-type, authorship, mode (e.g., language-based or not), and length.

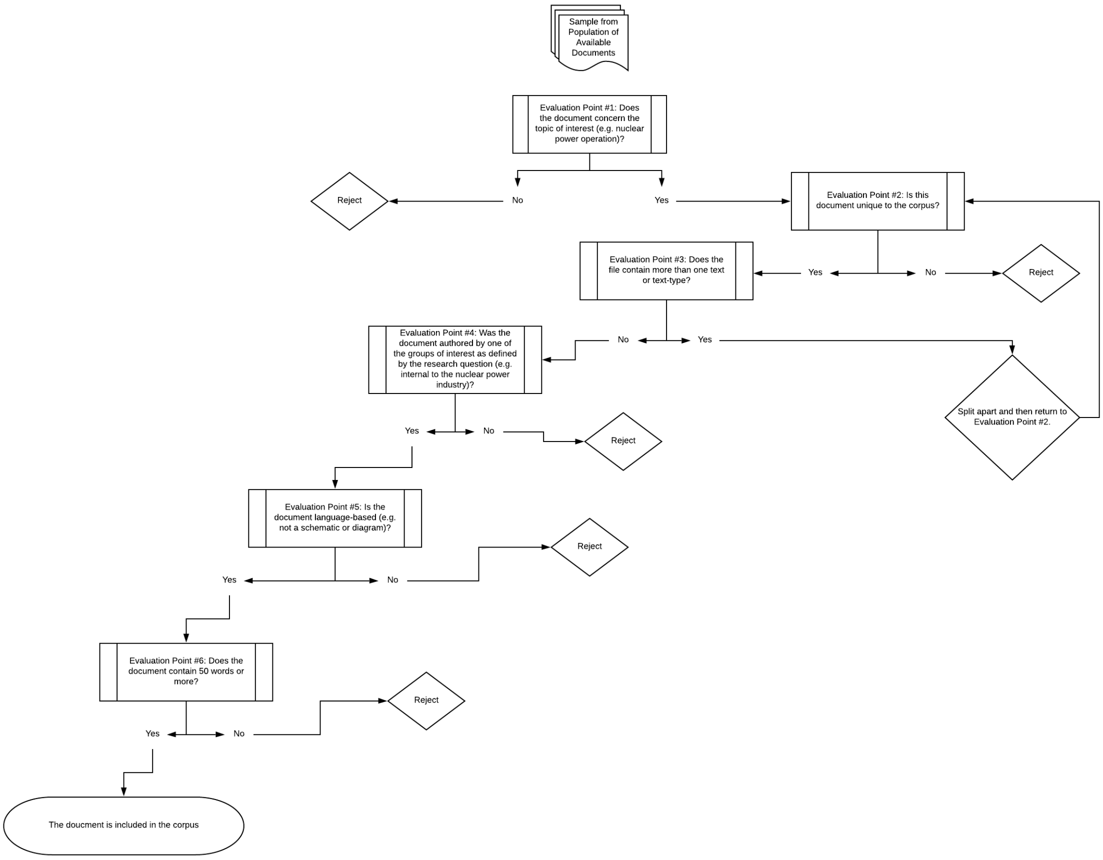

Sample from Population of Available Documents
Evaluation Point #1: Does the document concern the topic of interest (e.g. nuclear power operation)?
Reject
No
Yes
Evaluation Point #2: Is this document unique to the corpus?
Evaluation Point #3: Does the file contain more than one text or text-type?
Yes
No
Reject
Evaluation Point #4: Was the document authored by one of the groups of interest as defined by the research question (e.g. internal to the nuclear power industry)?
No
Yes
Split apart and then return to Evaluation Point #2.
Yes
No
Reject
Evaluation Point #5: Is the document language-based (e.g. not a schematic or diagram)?
Reject
Yes
No
Evaluation Point #6: Does the document contain 50 words or more?
Reject
Yes
No
The doucment is included in the corpus

Figure 2. Corpus sampling process diagram.

5. Results

One of the first observations made through the 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 manually 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 3120 documents downloaded from the ADAMS were reduced to 2775 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 3, there is a very distinct, and steep, asymptotic hyperbolic curve, or A-curve.

In the case of the data in Figure 3, 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, 1125 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 4773 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 4).

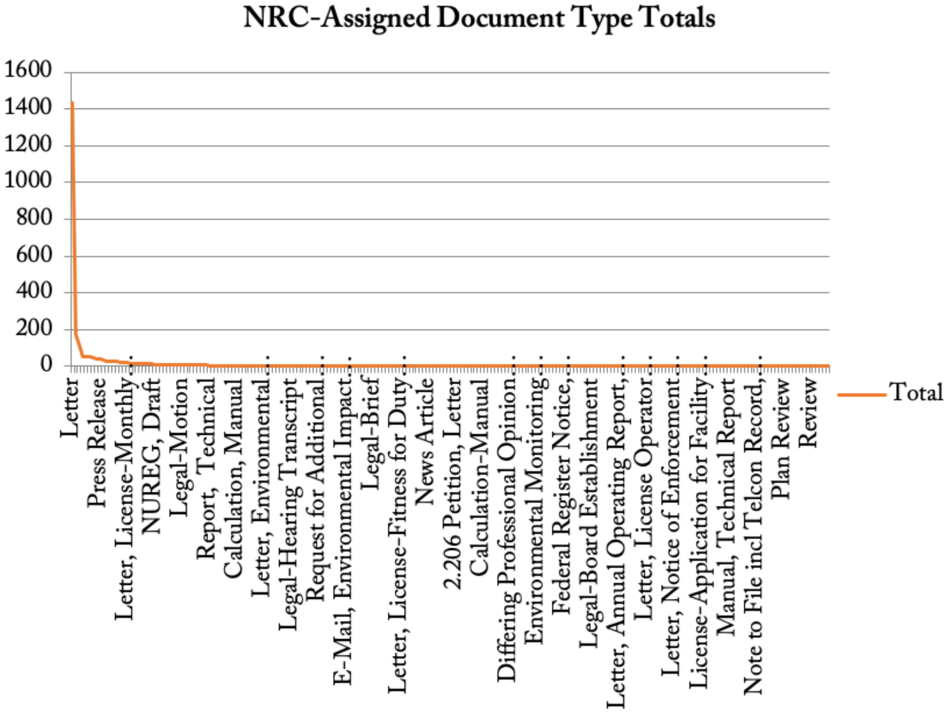


Figure 3. Pilot Document Totals Before Splitting Multiples.

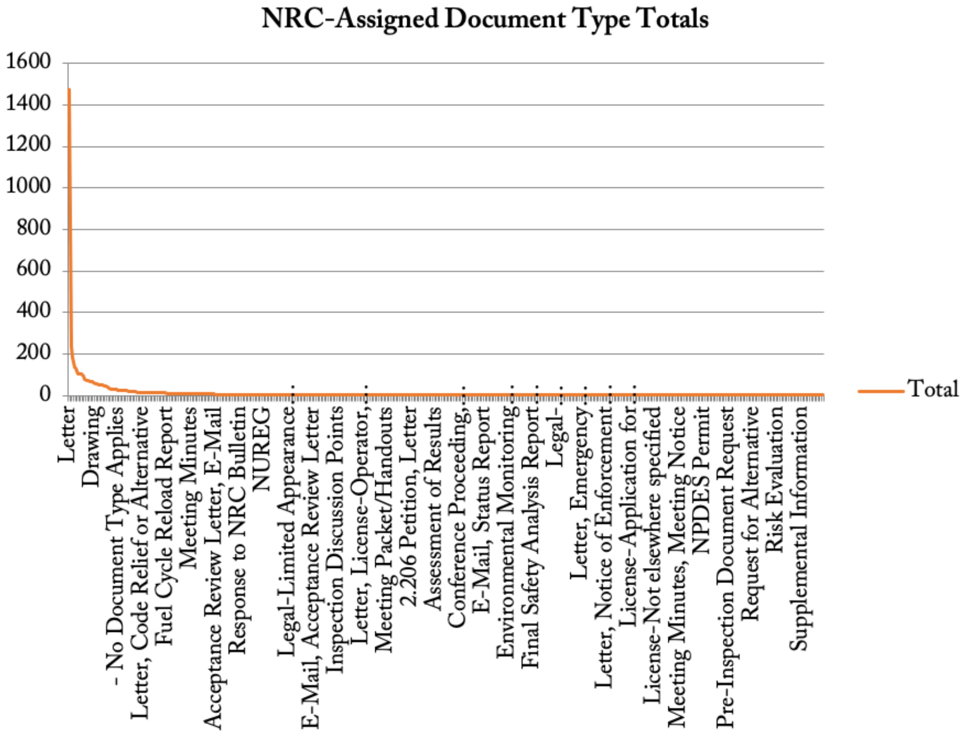


Figure 4. Pilot Document Totals After Splitting Multiples.

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 [20]. 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 4773 documents from the ADAMS-PARS database, 97.76% of them were authored by internal sources. Thus, 4666 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" [39], both internally and externally directed documents are maintained as part of the corpus.

Of the 4666 documents remaining in the Pilot, only 2.27% (or 106 of them) were not language-based documents, such as drawings and photographs (Figure 5). They were not kept as part of the reference corpus. For the 4560 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 2996 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 (3806) 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 3593 documents. In other words, the Pilot had a rejection rate of 24.72%. This rejection rate does not reflect the quality of the data, but rather the conformity of the population of texts to the question of interest. Essentially, 24.72% of the original sample were not adequate for inclusion in the text-based model for eventual analysis.

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 and then manually rated in order to look for consistency in the proportions of document rejection to create a sizable reference corpus from the ADAMS database. Only four iterations were conducted due to the resource-intensive nature of the process—as it relied on the visual assessment and verification of each document against the metadata generated by the database itself. Table 5 below provides the general statistical characteristics of each iteration after all evaluation was completed. Please note the highly variable nature of the size of the average document within the database that was sampled.

Table 5. General descriptive statistics for each iteration of principled subsampling.

Iteration	Min Word Count	Max Word Count	Average Word Count	Standard Deviation Word Count
Pilot	50	3465	930.68	1049.62
Iteration 2	50	3465	882.91	1009.18
Iteration 3	56	3320	929.37	1062.00
Iteration 4	50	3430	931.17	1040.15

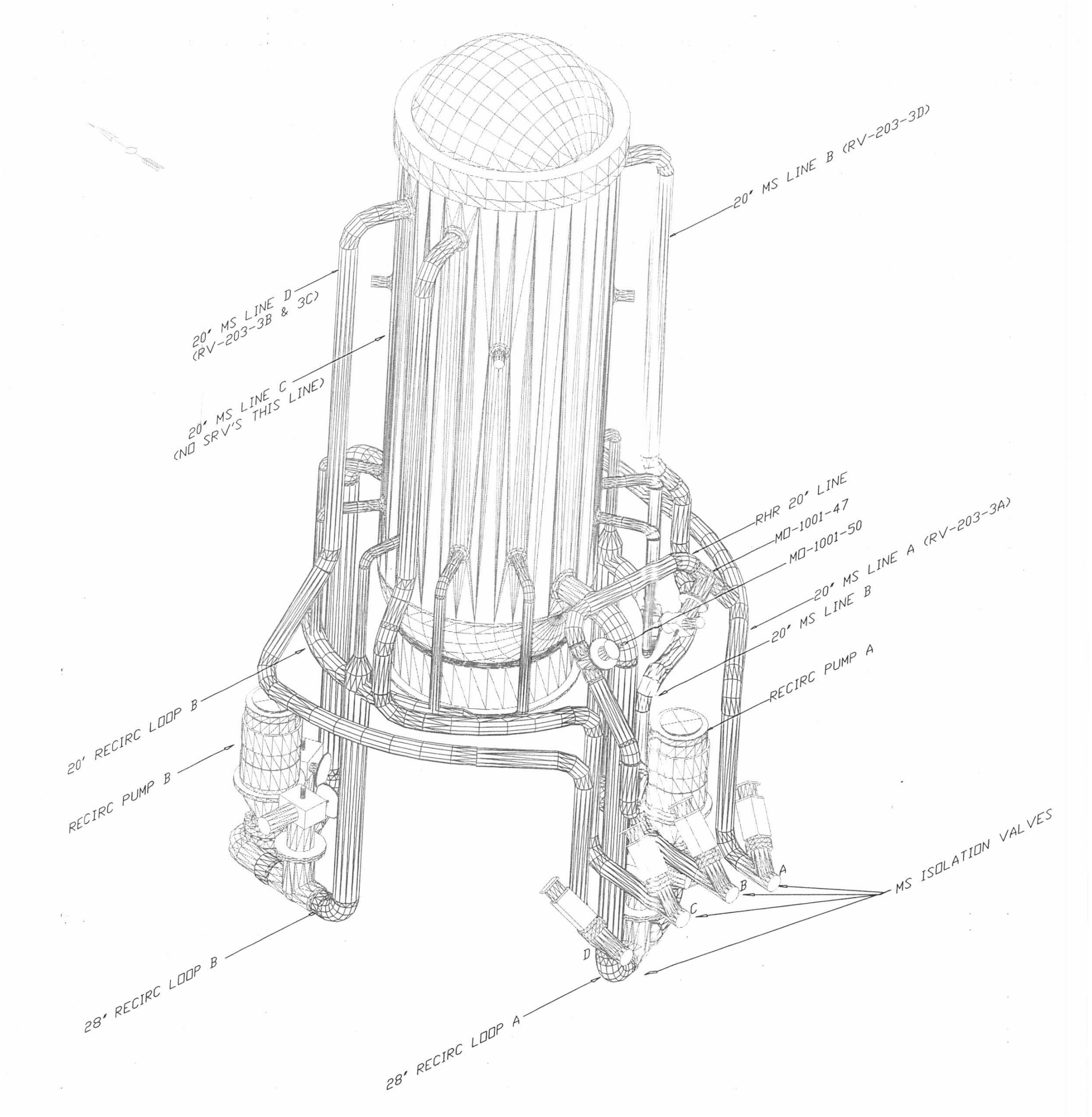


Figure 5. ML022530285 Drawing Handout from 7/24/2002 meeting regarding Pilgrim proposed safety-relief valve seismic analysis methodology.

Reproducibility

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.

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 6).

Table 6. Evaluation Point 2: Duplicate Accession ID Rejection Ratios.

Iteration	Total Documents	Number of Duplicate Documents	Rejection Ratio
Pilot	3120	345	11.06%
Iteration 2	3120	355	11.38%
Iteration 3	3120	368	11.79%
Iteration 4	3120	371	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 7).

Table 7. Evaluation Point 3: Ratio of Original Number to Number After Splitting Multiples.

Iteration	Number of Documents after Evaluation Point 2	Number of Documents after Splitting	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 8).

Table 8. Evaluation Point 4: Externally Authored Document Rejection Ratios.

Iterations	Number of Documents after Evaluation Point 3	Externally Authored 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 9).

Table 9. Evaluation Point 5: Non-Language-Based Document Rejection Ratios.

Iterations	Number of Documents after Evaluation Point 4	Non-Language-Based Documents	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 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 10), but also concerning the total rate of rejection for iterations two through four in comparison to the pilot sample (Table 11).

Table 10. Evaluation Point 6: Document Length Rejection Ratios.

Iterations	Number of Documents after Evaluation Point 5	Documents Having <50 Words	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 11. Total Rejection Ratios for All Iterations.

Iterations	Total Documents Sampled before Evaluation Point 6	All Documents Rejected	Rejection Ratio
Pilot	1180	4773	24.72%
Iteration 2	1110	4625	24%
Iteration 3	1059	4618	22.93%
Iteration 4	1040	4581	22.70%

The final text-based model, or corpus, of texts that resulted from the initial random sample that were then evaluated using the principle-driven schema led to a final mean rejection ratio of 23.5875 and variance of 0.891. This analysis provides an additional level of confidence that the sampling procedure pioneered for the Tobacco Documents corpus and 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 for a mutually exclusive database from a different domain. Such an assertion is predicated on the notion that such databases must have comprehensive metadata that document both linguistic and extralinguistic factors.

6. Discussion and Conclusions

The need for Big Language Data sampling approaches for constructing text-based corpora that adequately model and reflect the purpose of the analysis to be performed by a researcher is of critical importance to scientists and scholars interested in leveraging such methods to investigate issues of national or global importance. In the case of energy and social science research, it has been argued that there is a need for stronger, more creative approaches to research design and quality, as well as more innovative methods for investigating questions that are relevant and impactful for society (Sovacool et al. 2018). Specifically, their field has a need for methods that will assist them in investigating use-inspired research questions. The construction and analysis of Big Language Data corpora, like that demonstrated in this paper, have the potential to fill this need if they are sampled with principles that reflect the same extralinguistic factors that influence socially useful research.

By systematically analyzing and demonstrating the reproducibility and validity of the principled subsampling method first pioneered with the Tobacco Documents Corpus [34], such an approach can be used responsibly to investigate socially relevant issues that can be modeled with Big Language Data corpora. Search engines, like Google, and databases are messy, and we really must understand what they look like before subjecting them to analytical methods that may be at odds with their very organization. Machine-readable text artifacts of the human experience can be just as messy and complicated as the humans that make them. Much like performing an individual critical reading of a text, we as data scientists need to be careful and take the time to document, analyze, and understand how our readings of a database can be used as frameworks for creating subsampling plans for Big Language Data that can have real impact on society.

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 [44]. Moreover, the observation that this sampling approach is reproducible across multiple iterations now opens the possibility of future research in the use of machine learning and automated approaches for executing sampling frameworks for Big Language data and the assessment document compliance for inclusion: eliminating the need for costly human resources. There is also a distinct opportunity for future research around designing corpora from Big Language databases, knowledge systems, and other born digital sources that exhibit characteristics of complex systems. For socially relevant research using Big Language Data that facilitates investigations of social use, it should not be the focus of data scientists to reduce rejection rates but rather to develop automated methods for making such parsing processes more efficient for users who need to filter out the documents that do not fit their research questions that are motivated by non-linguistic characteristics. Extralinguistic factors and linguistic characteristics of documents sampled in the creation of corpora have the potential to be interconnected to a high degree and should be further investigated. Blending principled sampling frameworks with demographic sampling through human-centered design for corpus building 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.

Funding: This research received no external funding.

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*, 11 February 2012; pp. 1–5. [CrossRef]
2. Johansson, S. Times change, and so do corpora. In *English Corpus Linguistics*; Routledge: Abingdon, UK, 2014; pp. 305–314.
3. Johansson, S.; Stenström, A. *English Computer Corpora: Selected Papers and Research Guide*; Walter de Gruyter: Berlin, Germany, 1991; Volume 3.
4. Baker, P. *Using Corpora in Discourse Analysis*; A&C Black: London, UK, 2006.
5. Boyd, D.; Crawford, K. Six Provocations for big data. *SSRN Electron. J.* **2011**, *123*. [CrossRef]
6. Manovich, L. Trending: The promises and the challenges of big social data. In *Debates in the Digital Humanities*; University of Minnesota Press: Minneapolis, MN, USA, 2011; pp. 460–475.
7. Davies, M. TIME Magazine Corpus (100 Million Words, 1920s–2000s). 2007. Available online: http://corpus.byu.edu/time (accessed on 15 November 2018).
8. Kilgarriff, A.; Grefenstette, G. Introduction to the special issue on the web as corpus. *Comput. Linguist.* **2003**, *29*, 333–347. [CrossRef]
9. Introna, L.D.; Nissenbaum, H. Shaping the web: Why the politics of search engines matters. *Inf. Soc.* **2000**, *16*, 169–185.
10. Meyer, C.F.; Nelson, G. Data collection. In *The Handbook of English Linguistics*; Wiley-Blackwell: Hoboken, NJ, USA, 2006; pp. 36–93.
11. Kennedy, G. *An Introduction to Corpus Linguistics*; Longman: London, UK, 1998.
12. Sovacool, B.; Axsen, J.; Sorrell, S. Promoting novelty, rigor, and style in energy social science: Towards codes of practice for appropriate methods and research design. *Energy Res. Soc. Sci.* **2018**, *45*, 12–41. [CrossRef]
13. Blackwell, M.; Sen, M. Large datasets and you: A field guide. *Political Methodol.* **2012**, *20*, 2–5. Available online: http://www.mattblackwell.org/files/papers/bigdata.pdf (accessed on 15 November 2018).
14. Sinclair, J. *Corpus and Text: Basic Principles. Developing Linguistic Corpora: A Guide to Good Practice*; Wynne, M., Ed.; 2004. Available online: http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm (accessed on 29 December 2018).

15. Kretzschmar, W.A. *Language and Complex Systems*; Cambridge University Press: Cambridge, UK, 2015.
16. Kretzschmar, W.A. Language variation and complex systems. *Am. Speech* **2010**, *85*, 263–286. [CrossRef]
17. Burkette, A. The lion, the witch, and the armoire: Lexical variation in case furniture terms. *Am. Speech* **2009**, *84*, 315–339. [CrossRef]
18. McEnery, T.; Wilson, A. *Corpus Linguistics: An Introduction*; Edinburgh UP: Edinburgh, UK, 2001.
19. Stubbs, M. *Words and Phrases: Corpus Studies of Lexical Semantics*; Blackwell Publishing: Malden, MA, USA, 2002.
20. Kretzschmar, W.A., Jr. *The Linguistics of Speech*; Cambridge UP: Cambridge, UK, 2009.
21. McCarty, W. Modeling: A Study in words and meanings. In *A companion to Digital Humanities*; Schreibman, S., Siemens, R., Eds.; Blackwell: Malden, MA, USA, 2004; pp. 254–272.
22. Meyer, C.F. *English Corpus Linguistics: An Introduction*; Cambridge UP: Cambridge, UK, 2004.
23. Lohr, S. *Sampling: Design and Analysis*; Cengage Learning: Boston, MA, USA, 2009.
24. Kretzschmar, W.A.; Meyer, C.F.; Ingegneri, D. Uses of inferential statistics in corpus studies. *Lang. Comput.* **1997**, *20*, 167–178.
25. Anderson, W.; Corbett, J. *Exploring English with Online Corpora*; Macmillan International Higher Education: New York, NY, USA, 2017.
26. Biber, D. Representativeness in corpus design. *Lit. Linguist. Comput.* **1993**, *8*, 243–257. [CrossRef]
27. Biber, D. Methodological issues regarding corpus-based analyses of linguistic variation. *Lit. Linguist. Comput.* **1990**, *5*, 257–269. [CrossRef]
28. Biber, D. Using register-diversified corpora for general language studies. *Comput. Linguist.* **1993**, *19*, 219–241.
29. Gries, S.T. Dispersions and adjusted frequencies in corpora. *Int. J. Corpus Linguist.* **2008**, *13*, 403–437. [CrossRef]
30. McEnery, T.; Xiao, R.; Tono, Y. *Corpus-Based Language Studies: An Advanced Resource Book*; Routledge: New York, NY, USA, 2006.
31. Crowdy, S. Spoken corpus design. *Lit. Linguist. Comput.* **1993**, *8*, 259–265. [CrossRef]
32. Sampson, G. The empirical trend: Ten years on. *Int. J. Corpus Linguist.* **2013**, *18*, 281–289. [CrossRef]
33. Pechenick, E.A.; Danforth, C.M.; Dodds, P.S. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* **2015**, *10*, e0137041. [CrossRef] [PubMed]
34. Kretzschmar, W.A.; Darwin, C.; Brown, C.; Rubin, D.; Biber, D. Looking for the smoking gun: Principled sampling in creating the Tobacco Industry Documents Corpus. *J. Engl. Linguist.* **2004**, *32*, 31–47. [CrossRef]
35. Kretzschmar, W.A. *Sampling Plan for Creation of Corpora for the Tobacco Documents Grant*; Self published: Athens, GA, USA, 2001.
36. Mudraya, O. Engineering English: A lexical frequency instructional model. *Engl. Spec. Purp.* **2006**, *25*, 235–256. [CrossRef]
37. Walker, S.J.; Wellock, T.R. *A Short History of Nuclear Regulation, 1946–2009*; U.S. Nuclear Regulatory Commission: Rockville, MD, USA, 2010.
38. Bodansky, D. *Nuclear Energy: Principles, Practices, and Prospects*; Springer: New York, NY, USA, 2004.
39. The United States Nuclear Regulatory Commission. *About NRC*; The United States Nuclear Regulatory Commission: Rockville, MD, USA, 2018. Available online: https://www.nrc.gov/about-nrc.html (accessed on 15 November 2018).
40. Henry, C.L. *Freedom of Information Act*; Nova Publishers: Hauppauge, NY, USA, 2003.
41. The United States Nuclear Regulatory Commission. *Withholding of Sensitive Information for Nuclear Power Reactors*; The United States Nuclear Regulatory Commission: Rockville, MD, USA, 2018. Available online: http://www.nrc.gov/reading-rm/sensitive-info/reactors.html (accessed on 15 November 2018).
42. The United States Nuclear Regulatory Commission. *ADAMS Public Documents*; The United States Nuclear Regulatory Commission: Rockville, MD, USA, 2018. Available online: http://www.nrc.gov/reading-rm/adams.html (accessed on 15 November 2018).

43. Hettel, J. Harnessing the Power of Context: A Corpus-Based Analysis of Variation in the Language of the Regulated Nuclear Industry. Ph.D. Thesis, University of Georgia, Athens, GA, USA, 2013. Available online: https://getd.libs.uga.edu/pdfs/hettel_jacqueline_m_201305_phd.pdf (accessed on 15 November 2018).
44. Březina, V.; Meyerhoff, M. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *Int. J. Corpus Linguist.* **2014**, *19*, 1–28. [CrossRef]