The Extent of Repetition in Contract Language

Dan Simonson
BlackBoiler LLC
Arlington, VA 22207
dan.simonson@blackboiler.com

Daniel Broderick
BlackBoiler LLC
Arlington, VA 22207
dan@blackboiler.com

Jonathan Herr
BlackBoiler LLC
Arlington, VA 22207
jonathan@blackboiler.com

Abstract

Contract language is repetitive (Anderson and Manns, 2017), but so is all language (Zipf, 1949). In this paper, we measure the extent to which contract language in English is repetitive compared with the language of other English language corpora. Contracts have much smaller vocabulary sizes compared with similarly sized non-contract corpora across multiple contract types, contain $1/5^{th}$ as many hapax legomena, pattern differently on a log-log plot, use fewer pronouns, and contain sentences that are about 20% more similar to one another than in other corpora. These suggest that the study of contracts in natural language processing controls for some linguistic phenomena and allows for more in depth study of others.

1 Introduction

Among attorneys and those in the legal professions, contract language is considered "repetitive," but the same can be said about natural language in general (Zipf, 1949). Anderson and Manns (2017) largely attribute the repetitive nature of contract language to drafting methodologies. Attorneys rarely start contracts or provisions from a blank document. Anderson and Manns (2017) showed that attorneys typically select a precedent contracts and fit them to the parameters of a new relationship with a counterparty. The same is true when an attorney drafts a new contract provision, beginning with an old provision from an existing agreement. As a result, new or novel language is infrequent as compared to other kinds of natural language. They assessed these similarities using Levenshtein distance, which is somewhat unusual with respect to the methods and statistics typically used in natural language processing and corpus linguistics, and they did not compare their corpus of contracts against data of the sort typically used in natural language processing and corpus linguistics—for our purposes, the Brown Corpus (Francis and Kučera, 1964, 1971, 1979) and Wikipedia (King, 2018). This paper seeks to describe quantitatively the extent to which contract language is more repetitive than the language found in these corpora.

We aim this paper at multiple audiences. Our own motivation was to more deeply understand the driving linguistic and distributional factors behind technology that Broderick et al. (2016) developed, and we hope those in industry who work with or are evaluating legal technology can read this work to understand how this repetition uniquely supports the automation of work involving contract language. We hope the computational linguist working on contract or legal texts can use the findings here to justify certain decisions and positions made in their own work, as a basic foundation of facts can help reduce exponentially the tree of decisions made in practice. We hope the computational linguistics community at large can take from this paper that contract language has properties advantageous for problems that prefer more constrained—but still natural—language.

In this paper, we present analyses of contract language in English, juxtaposing them against two other English language corpora. In Section (2), we discuss prior work toward this end. In Section (3), we discuss the data used in this study: a set of contract corpora containing 1,737 documents and two baseline corpora for comparison. In Section (4), we discuss the distribution of tokens in our contract corpus compared with the baseline corpora and look more closely at the data to affirm the meaning of those distributions. In Section (5), we step up from the token level to look at similarity at the sentential level through a nearest neighbors analysis. In Sections (6 & 7), we discuss these findings broadly and conclude.

Proceedings of the Natural Legal Language Processing Workshop 2019, pages 21–30
Minneapolis, Minesotta, June 7, 2019.

2 Prior Work

Numerous studies have been done on contract corpora, many with specific ends in mind. Faber and Lauridsen (1991) prepared a corpus referred to as the "Danish-English-French corpus in contract law," or the "Aarhus corpus" (Curtotti and McCreath, 2011). This corpus contains both contracts and legal literature, sampling what language was available at the time for study, and containing 776 pages of contracts, likely around 400,000 tokens.[1] A few corpus studies have been conducted previously. Blom and Trosborg (1992) look at speech acts in the Danish-English-French corpus and how different speech acts create the relationship established between two parties in a contract, as well as other types of legal language. Nielsen and Wichmann (1994) look at the English subcorpus of the Danish-English-French corpus of a size of around 50,000 tokens in conjunction with an equally sized corpus of their own creation in German, focused primarily on how obligation is expressed in both languages. Anesa (2007) examine a corpus of 12 contracts (50,828 tokens) in English to codify the linguistic strategies used to make certain provisions vague while others specific, particularly because contracts are rarely written from scratch, and vague provisions have broad applicability when cut-and-paste between contracts. Carvalho (2008) and Mohammad et al. (2010) both built parallel corpora for the purposes of improving contract translation between English and Brazilian Portuguese and Arabic respectively. Most extensively, Curtotti and McCreath (2011) conducted a study of a corpus of Australian English contracts, examining the distribution and statistics of a corpus of 256 contracts. They primarily focused on demonstrating that the corpus they had collected was representative, presenting numerous statistics to that end. Anderson and Manns (2017) collected a corpus of 12,407 agreements with the end of showing precedent relationships using Levenshtein distance and clustering. Their goal was not to analyze the properties of contract language itself, but to show explicitly how contracts are copied from one another.

Studies have also been done on repetitive, conventionalized language use more broadly. Halliday (1988) examined the historical development of conventionalized language use in the sciences, demonstrating qualitatively that science developed a rhetorical style using already existing rhetorical elements of English in a way most relevant to the experimental style of the physical sciences. Extensive corpus work has looked at the physical sciences to better understand the linguistic processes of used to create and frame understanding in scientific work (Argamon et al., 2005) and to exploit the repetitive nature of such documents to find phrases that are more information-laden than their more conventionalized counterparts (Degaetano-Ortlieb and Teich, 2017).

3 Data

In this section, we describe the data used in this study, both the contracts we gathered for analysis and the baseline corpora used to contrast the contract corpus.

3.1 Contract Corpus

Our subject matter expert gathered a corpus of contracts. They selected categories and recovered documents relevant to those categories through search engines and from EDGAR,[2] issuing queries based on phrases indicative of particular contract types, and selecting contracts that they considered, as an expert, to be of the specific contract type.

Our subject matter expert searched for five types of contracts. Prime Contracts are an agreement between a project owner and a general contractor. Subcontracts are between a general contractor and a subcontractor or trade contractor for specific subcomponents of a project, such as implementing the drywall in the building. Non-Disclosure Agreements are agreements related to the exchange of information and the confidential treatment thereof. Purchase Orders is an agreement for the purchase of products between a buyer and a seller. Services Agreements is an agreement for one party to supply another party with a service. Prime Contracts and Subcontracts were selected with relation to the construction industry; the other types were selected more broadly.

Contracts identified with formatting issues due to failed optical character recognition (OCR) were removed. This filtering process was not perfect, however, as some OCR errors remain in the data.

[1] We were not able to obtain access to this corpus, and are estimating this based on the provided page count (Faber and Lauridsen, 1991) at 500 tokens per page.

[2] EDGAR is a service of the United States Securities and Exchange Commission: (`https://www.sec.gov/edgar.shtml`)

To some extent, we have to acknowledge that some noise in contract data is inevitable; there are no universal standards for the interchange of contract documents. Further, while the contracts retrieved represent unique agreements, they themselves are not necessarily unique. This is intended to represent the typical distribution and content of contract documents exchanged on a regular basis. Table (1) contains the results of this retrieval process, a total of 1,737 documents containing a total of over 15 million tokens.

To find tokens and sentence boundaries, corpora were pre-processed with SpaCy (Honnibal and Johnson, 2015).[3]

Table 1: Document and Token Counts per Category.

Category	# docs	# tokens	toks/doc
NDAs	791	1,955,522	2,472
Prime Contracts	174	5,417,987	31,138
Purchase Order	229	1,933,547	8,443
Service Agreements	137	1,216,724	8,881
Subcontracts	406	5,029,433	12,388
Total	1,737	15,553,213	8,954

3.2 Baseline Corpora

We compare the contract corpora against two others: Wikipedia and the Brown Corpus.

For Wikipedia, we used King (2018)'s release of the encyclopedia on Kaggle. This was pre-processed with most mark-up removed and put into a SQLite database for easy accessibility and reproducibility, containing 4,902,648 articles. While the database is organized on a section-by-section basis, we retrieve all text article-by-article to mirror how we access contract documents.

The Brown Corpus (Brown) is a representative corpus of American English (Francis and Kučera, 1964, 1971, 1979). As the first digital corpus of natural language, the Brown Corpus has acted as a common litmus test across experimental configurations for decades. While more contemporary corpora exist (Mair, 1997; Davies, 2010), Brown is much easier to access, and while specific lexical items themselves may have changed over the last sixty years, we do not anticipate that the relative distribution of lexical items in the English language itself to have changed dramatically in that time. "fileids" are considered document boundaries; these are in some cases subsamples of whole documents. Brown was used through the interface provided by NLTK (Bird et al., 2009).[4]

To find tokens and sentence boundaries, corpora were pre-processed with SpaCy (Honnibal and Johnson, 2015).[5] This includes the Brown Corpus, which was re-tokenized to be consistent with the other corpora.

4 Rank-Counts Analysis

In this section, we present an analysis of the rank-vs-counts curves of the corpora analyzed. Often referred to as a *Zipfian* analysis, we counted the frequency of each token type and arranged the counts by their respective ranks—that is, the most frequent word has a rank of 1, the second most frequent word has a rank of 2. Natural language corpora approximately exhibit *Zipf's Law* (Zipf, 1949; Manning and Schütze, 1999)—that a word's frequency is inversely proportional to its rank ($f \propto r^{-1}$). This distribution is difficult to observe on a linear plot and is better observed on a log-log plot, where it appears linear.

For maximum comparability, we subsampled our corpora to the number of tokens in the smallest corpus, rounded up to the nearest document. In other words, we included one whole document at a time until we were just in excess of the number of tokens in the smallest corpus—in this case, the Brown Corpus with 1,161,192 tokens.

4.1 Results

Figure (1) contains a log-log plot of the rank-vs-counts for each word type in the subcorpora counted. Notably, the rank-counts curve of the contract corpus bends downward around rank 1000. This bend was also previously identified in Curtotti and McCreath (2011), who chose to justify that this deviation is normal for typical English language corpora. This is true, but juxtaposing the curves for the contracts and baseline corpora reveals that the curves of the contract subcorpora are far steeper. This means that there are far fewer rare word types in the contract corpora than in the baseline corpora—in other words, that rare words appear less often in contracts.

Inspection of the statistics describing the curves is more revealing. Table (2)[6] presents these val-

[3] https://spacy.io/

[4] https://www.nltk.org/data.html

[5] https://spacy.io/

[6] We included TTR largely as a matter of convention and since the number of tokens in each sampled subcorpus is similar.

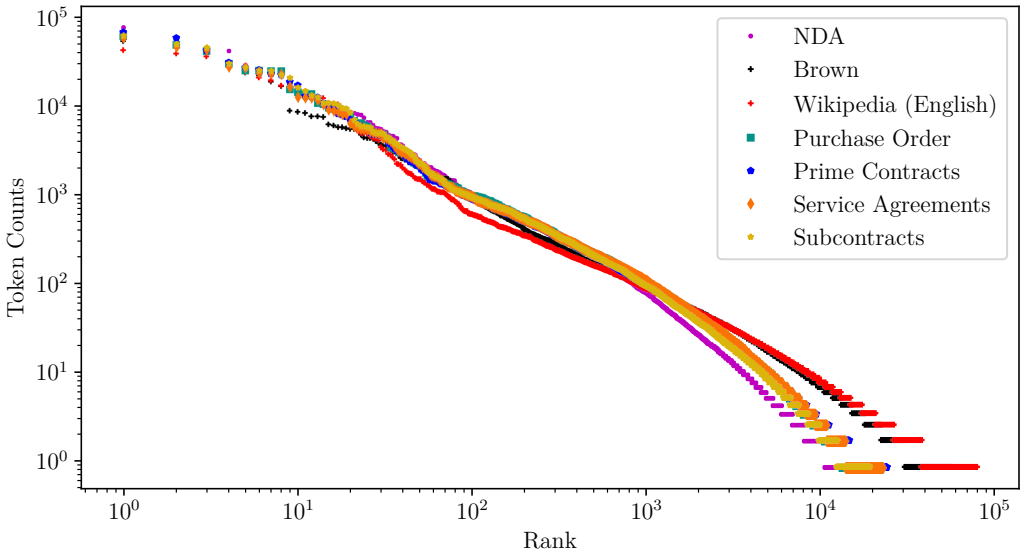

Figure 1: Log-log plots of rank vs frequency of tokens in each contract subcorpus and the baseline corpora.

Table 2: Raw Statistics on Subsamples of Corpora Investigated. $|C|$ indicates the size of the whole corpus; $|S|$ indicates the size of the subcorpus investigated.

Series	$\|C\|$	$\|S\|$	# Tokens	# Types	TTR	# Hapax	H/Types	H/Tokens
Brown	500	500	1,161,192	56,057	4.83%	25,559	45.59%	2.20%
Wikipedia:EN	4.9M	1,559	1,161,264	78,973	6.80%	40,820	51.69%	3.52%
NDA	791	484	1,164,051	17,454	1.50%	6,837	39.17%	0.59%
Purchase Order	229	132	1,164,421	21,670	1.86%	8,404	38.78%	0.72%
Prime Contracts	174	36	1,162,939	23,971	2.06%	9,461	39.47%	0.81%
Services Agreements	137	131	1,164,687	22,854	1.96%	8,915	39.01%	0.77%
Subcontracts	406	96	1,163,421	19,052	1.64%	6,670	35.01%	0.57%

ues, which exhibit a clear distinction between the baseline corpora and the contract corpora by all measures. The number of word types is more than double at the least extreme, Prime Contracts vs Brown, and is pentupled at the extreme, NDAs vs Wikipedia. These differences make sense. Prime Contracts are the most likely to be negotiated and tailored to the specific deal while NDAs are the least likely to be nitpicked, themselves often a preliminary step to generating the larger, revenue-generating deal.

The extreme case of a rare item is what is commonly referred to as a *hapax legomena*: a word that is only contained once in a corpus and never seen again. These are visible in Figure (1) toward the bottom right—the final "stair step" as the counts approach zero. Another measure to further illuminate the difference between the baseline and contract corpora is the ratio between the number of hapax legomena and the number of tokens in the corpus. Hapax / Tokens tells us "out of all the words we see in the corpus, how often do we encounter a word that we have never seen before and never see again." As fractions, for Brown, this ratio is 1/35; that is, every 35^{th} word we see once, and never again. For Wikipedia, it is about 1/24.

For NDAs, the Hapax / Tokens ratio is 1/135; for Prime Contracts, the ratio is around 1/120. Between the Prime Contracts and Wikipedia, this difference is nearly 1/5—that is, for every 5 hapax legomena in the Wikipedia corpus, there is 1 in the Prime Contracts corpus. Extremely rare word types—hapax legomena—do not appear in contracts as often as in Brown or Wikipedia.

4.2 Qualitative Inspection

This section presents the content of the corpora studied to validate, from a qualitative perspective, the patterns identified quantitatively. Specifically, we examine both of these sorts of items to give qualitative context to the statistics given in Section (4.1), understanding what these differences mean for a rare, open class of types (hapax legomena, Section 4.2.1) and a frequent, closed class of types (pronouns, Section 4.2.2).

4.2.1 Hapax Legomena

As discussed in Section (4), the corpora diverge with respect to the frequency of rare tokens in the corpora, particularly with respect to the number of hapax legomena. To further illuminate the nature of these, Table (3) contains examples of hapax legomena from each subcorpus.

Based on the given samples, we can see that in both cases, the hapax legomena are often numbers or proper nouns. The contract corpora are also more susceptible to generally noisy data. The use of capitalization for emphasis differentiated a lot of terms that otherwise appear frequently (e.g. "INTRODUCTION"), as well as unique numbering schemes (e.g. "FP-1"). Additionally, though despite our efforts to remove data in which OCR failed, a few examples slipped through of (e.g. "wri+en" and "totheChangeinwriting").

Given these differences, we suspect further improvements, such as using lemmas and more carefully removing OCR errors, would actually amplify the difference in numbers of hapax legomena between corpora. Wikipedia's are mostly proper nouns, so these would remain hapax legomena—even lemmatized—while many in the contract subcorpora would be lemmatized into another word or removed from the data entirely. These modifications will amplify whatever differences were observed in this experimental configuration.

4.2.2 Frequency and Use of Pronouns

In some ways, pronouns—words such as "she," "their," and "itself"—are the reverse of hapax legomena, often being amongst the most common words in a corpus. However, they too appear less often in contracts. In both baseline corpora, 73,521 pronouns appeared, while in the contract corpora, a combined 52,764 pronouns appeared despite being 2.5 times larger than the combined baseline corpora. This is presumably to achieve precision and reduce ambiguity, as pronouns are often ambiguous and must be determined from context, but not all pronouns pattern alike.

The log frequency ratio shows these differences quite well. We define the log frequency ratio in this case to be $LF_{a,b}(w) = \log_{10} \frac{f_a(w)}{f_b(w)}$, where a and b are corpora, w is a token type, and $f_a(w)$ is the frequency of token type w in corpus a. Intuitively, if $LF_{a,b}(w)$ is zero, w appears with equal frequency in a and b; if it is negative, w appears more often in b. With log base 10, $LF_{a,b}(w) = 1.0$ means w appeared in corpus a 10 times as often as in b, etc. We will refer to all contract corpora as C_c and all baseline corpora as C_b. *forms of x* include the nominative, accusative, reflexive, and possessive forms of the pronouns—so *forms of "they"* include "they," "them," "their," and "themselves."

Forms of "he" and "she" appear the most com-

Table 3: Examples of Hapax Legomena From the Subcorpora Analyzed.

Corpus	*Sample of Hapax Legomena Tokens*
Brown	'ARF', 'Piraeus', 'flint', 'Volta', 'paterollers', 'Schmalma', 'melanderi', 'bongo', 'hard-to-get', 'Beloved', 'miniscule', 'Tower', 'temerity', 'Fay', 'avidly', ...
Wikipedia:EN	'appropriates', 'Puschmann', 'Muin', 'AC.7', 'sensing', 'Ambas', 'Kalutara', 'Arnott', 'Ogrskem', '48/73', 'Jayan', 'MK2020', 'beauticians', ...
NDA	'disapprove', 'mostly', 'wri+en', '15260', '48104', 'Loving', 'EXCLUSIVE', 'Culver', 'Chih', 'Hwa', 'inch', 'Behalf', 'Opinions', 'HD8', 'appropriated', ...
Purchase Order	'ASNs', 'FRED', 'Party(i)wherethereceivingPartyistheSupplier' 'overturn', 'Navigation', 'work.(iii', 'PLU', 'CDI', 'DFFRUGDQFH', 'INFRINGE', ...
Prime Contracts	'executers', 'Quote', 'derrick', 'FP-1', 'FP-3', 'FP-2', '00:00', 'Ceiling', 'EQUITABLE', 'OBTAIN', 'fan', 'ticket', 'prolonged', 'Macao', '19.2.2(c', ...
Services Agreements	'Sophia(R', 'sacrifice', 'adhesives', 'transloader', 'totheChangeinwriting', 'salient', 'simulate', 'KG', '15/29', 'divert', 'ownedbyCorsearch', 'biologic', ...
Subcontracts	'ENCOURAGED', 'closer', 'INTRODUCTION', 'projecting', '14607', 'CUT', 'Higher', 'interfaces', 'percipient', 'takeover', 'postponement', 'timesheet', ...

paratively infrequently, with a $LF_{C_c,C_b} = -1.43$ Typically, these are anaphoric; the referent is in the exact sentence where the lexical item was used, e.g: "*Employee* agrees that all information communicated to *him/her* concerning the work..." where the referrent is contained in the same sentence as the pronoun. Use of these pronouns are comparatively rare in contracts.

Forms of "they" and "we" appear comparatively infreqently as well, with an $LF_{C_c,C_b} = -0.64$ and $LF_{C_c,C_b} = -0.54$ respectively. Contracts occasionally will use "we" to denote one of the parties involved in the agreement. Similarly, forms of "you" appear quite often compared to other pronouns, with $LF_{C_c,C_b} = -0.12$. In fact, compared against Wikipedia, the $LF_{C_c,\text{Wikipedia}} = 0.57$, which means contracts use "you" far more often than it is found in Wikipedia. "you" simply does not make sense in an encyclopedic style, while on the other hand, contracts will use "you" in a similar manner to "we," defining who exactly "you" refers to at the beginning and never changing throughout, for example, "...the terms "*you* and *your*" are used in this Agreement, the same shall be construed as including..." defines exactly who "you" refers to. Thus, because "you" always refers to the same entity or set of entities throughout a document and because its deictic quality makes it unambiguous, this allows for its occasional use in contracts.

Last of all, forms of "it" appeared with an $LF_{C_c,C_b} = -0.11$. Like with the rare uses of "he" or "she," "it" is also used in tightly constrained contexts where the referent appears in close proximity to the pronoun it refers to, e.g: "Each *Contract Party* warrants that *it* has the right to make disclosures..." has both the referent and the pronoun in the same sentence, like "he/him" above. Similarly, the cataphoric use remains as well, e.g: "...*it* is *the intention of the Recipient* to give the Information Provider the broadest possible protection..." has the same local quality as the anaphoric use of "it." Nevertheless, forms of "it" appear less often than in the baseline corpora. It is often preferred to using "he," "she," or "they" to refer to one of the defined parties in a contract, though it is still used 23% less often than expected.

As we can see, pronouns such as "she," "he," and "they" appear far less often than in other English corpora; "it" is often used instead, parties are referred to with the pronouns "we" and "you," or the explicit name of the party is used. Even so, these pronouns appear with less frequently than expected. This makes problems like anaphora resolution—one of the more sophisticated components of a coreference resolution system and quite a challenge typically—less difficult to perform on contract language.

5 Nearest Neighbors

Section (4) showed clear differences in the distribution of tokens between contract corpora and the baseline corpora. But tokens alone do not entail repetition, especially if contracts are using the

same tokens but in novel ways. One technique to address this question is to compare the sentences within a corpus—to find, for every sentence in the corpus, what the next most similar sentence is. This gives some idea of how repetitive a corpus is at the sentential level.

We considered an information theoretic approach to this problem (Shannon, 1948; Harris, 2002; Crocker et al., 2016; Degaetano-Ortlieb and Teich, 2018). However, we wanted results that were as accessible as possible to a more general audience, and information theoretic results require some initiation in the nuances and meanings of bits and entropy for their meaning to be clear. Instead, we primarily use a unigram vector model (Manning and Schütze, 1999, Section 8.5.1). In a unigram vector model, values derive from an intuitive pairing of identical tokens. Additionally, since the range of possible values falls on a scope from zero to one, their reflection of the similarity between sentences is clear as well, with 0.0 indicating no shared tokens and 1.0 meaning that all tokens between two sentences are shared.

In this section, we conduct a second analysis of the corpora, in this case on a sentence-by-sentence basis, as opposed to merely looking at the rankings individual tokens. Our goal is to see, as we add documents from the each corpus to a collection of sentences under consideration, how the distribution of similarities between sentences changes. This will validate whether contract documents are more repetitive at the sentential level.

5.1 Vector Model

From a sequence of tokens, we define a vector v:

$$\vec{v} = w_0\vec{w}_0 + \ldots + w_i\vec{w}_i + \ldots + w_n\vec{w}_n \quad (1)$$

where w_i is the weight of the vector in dimension $\vec{w}_i$, and n is the vocabulary size in the subcorpus under analysis. Each vector dimension corresponds with a token type in the corresponding sentence. We normalize the weights of each vector dimension such that:

$$w_i = \frac{c_i}{\sqrt{\sum_j c_j^2}} \quad (2)$$

where c_i is the counts of token type i in the sentence under analysis. We compute the cosine similarity between two vectors $\vec{v}$ and $\vec{v}'$ with the dot product:

$$\vec{v} \cdot \vec{v}' = \sum_i w_i \times w_i' \quad (3)$$

which is, due to the weighting, valued between 0 and 1.

From each document, vectors are prepared for each sentence. If a sequence is shorter than 5 tokens, it is removed from consideration. Adding one document at a time, and for each sentence in each subcorpus, we compute the highest dot product value with any other sentence in the corpus, excluding itself. This is done by computing all dot products and storing the highest value.

Because of the sheer number of calculations, this is quite a computationally intensive experiment. We restrict our analysis to the NDA corpus vs Brown and Wikipedia. Because the number of comparisons grows exponentially, we further restrict our analysis to the first 200 randomly selected (without replacement) documents for each corpus.

5.2 Results

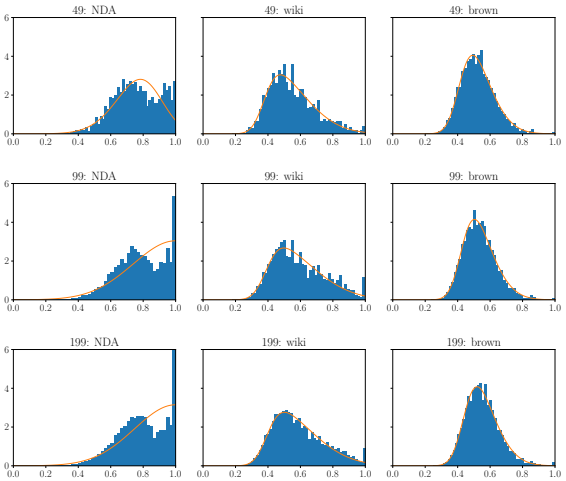

Figure 2: Histograms of the Distribution of Some of the Scores in the Nearest Neighbors Analysis. The histogram bins have a combined area of 1.

After filtering for sequences shorter than 5 tokens, the 200 document subsection of the NDA corpus contains 11,765 sentences; the Brown subsection contains 20,396; the Wikipedia subsection contains 5,385.

Scores generated at some substeps are featured in Figure (2). For the distribution of scores at each step, we fit a skewed normal distribution (Azzalini and Capitanio, 1999), as implemented in SciPy.[7] All fits were statistically significant ($p < 0.001$) past 22 documents. Figure (2) shows histograms

[7] https://www.scipy.org/

of the distribution of scores at three samples of these steps at 50, 100, and 200 documents.

Table 4: Statistics of the Distribution of Scores by Number of Docs. "Average" refers to the average of all scores at that point. "Frac Max" refers to the fraction of scores in the highest bin between 0.98 and 1.

Corpus	Statistic	@ 50	@ 100	@ 200
NDA	Average	0.760	0.783	0.791
Wiki	Average	0.551	0.592	0.590
Brown	Average	0.524	0.536	0.552
NDA	Frac Max	4.69%	10.65%	12.15%
Wiki	Frac Max	0.89%	2.03%	1.58%
Brown	Frac Max	0.11%	0.15%	0.26%

Over all, as documents are added, the average scores slowly trend upward (Table 4) between three and four percent. The number of sentences that are near exact matches ("Frac Max") increases dramatically in the NDA corpus compared with the baseline corpora as documents are added, with almost 8% of the scores moving to the near-exact match bin. With respect to the baseline corpora, this barely moved or decreased slightly.

5.3 Discussion

The results indicate quite clearly that repetition is even more salient at the sentential level than at the token level. While we saw trends in the tokens that showed a greater extent of repetition in the contract corpora at large, it is clear this holds at the sentential level, at least for NDAs, with an average sentence similarity that is 20% higher over all. Further, the number of sentences that are identical at 200 documents is almost 11% greater in the NDA data than in the baseline corpora.

Looking closely at the exact matches—sentences with a neighbor from 0.98 to 1.0—many in the NDA corpus are identically worded sentences. However, they are not identical enough that exact string matching is sufficient—for example, they contain the same word types with different capitalization and punctuation. There are, of course, examples of exact repetition in the baseline corpora too. Wikipedia articles contain infoboxes that appear in multiple articles, and parts of them parsed as sentences are exact matches. Brown newswire contains datelines, some of which are the same between articles: e.g. "Miami, Fla., March 17 –". The repetition of exact phrases is not a linguistic phenomenon unique to contracts; however, its frequency is greatly increased.

Most sentences in all corpora do not have an exact match. The average gives us some idea of the kind of match a typical sentence can make. Cosine distance at the sentential level is quite picky. These two pairs of sentences both have the average similarity of around 0.79:

(1a) Any assignment without such written consent shall be null and void and of no force or effect.

(1b) Any such attempted assignment shall be void and of no effect.

(2a) Notwithstanding any other provision of this Agreement to the contrary, this Agreement shall be effective as of the date first above written and shall remain in full force and effect thereafter for a period of two (2) years, whereupon the Agreement shall automatically terminate, unless otherwise terminated by the mutual written agreement of the Parties.

(2b) This Agreement shall be effective as of the Effective Date and continue for a period of five (5) years, or until termination of the Relationship, unless this Agreement is earlier terminated by mutual written agreement of the parties.

We can see clear similarities between the pairs of sentences. Both perform roughly the same function in a contract, albeit with some variation in wording and specific parameters. This shows that even a simple alignment technique can provide quality alignments between similar sentences across contracts, and given that this is the average similarity, that contracts indeed share quite a lot in common with one another, even when that is not an exact match between the two.

Examining the distributions of each of the steps, clearly both the Brown and Wikipedia corpora are modeled quite well by the skew normal distributions, with the curves clearly following the histograms they describe. The skew distribution of the NDA corpus also fits significantly, though the fit does not so obviously model the data; the sheer number of counts contained in the data allowed for the model to be significant. It may be better modeled by a superposition of two distributions, one covering near exact matches and the other covering the broader distribution. Such a model may parametrically work to model the exact matches

in the other two baseline corpora too, resulting in improved modeling for those as well.

Regardless of how we judge these models, the degree of repetition in NDAs is much greater than the two baseline corpora, with a great quantity of scores skewed toward 1.0, and even the secondary peak appearing just below 0.8, while the two baseline corpora peak between 0.5 and 0.6. By any measure, the NDA corpus is far more repetitive at the sentential level than its baseline counterparts.

6 Discussion

Contract language, as opposed to most natural language corpora typically studied, is far less variable, exhibiting far fewer rare word types and a much higher sentence similarity. These meet our expectations, and we can definitively state the extent to which contract language is repetitive: hapax legomena appear with $1/5^{th}$ the frequency of other corpora and sentences are 20% more similar on average. These are big differences, but not so big that they defy expectation.

One may be tempted to take this to the extreme and claim that the content of contracts is itself not language at all,[8] but this is a slippery slope fallacy. There is middle ground between newswire and non-language; while contract language may be more repetitive, this does not entail that it is not language. In fact, repetition is a core element of all language. With respect to tokens, this is embodied in the quantitative Zipf's law, but even at the discourse level, notions of repetition like *intertextuality* facilitate what we say and why we say it (Kristeva, 1980). Idioms like "I pronounce you husband and wife" have been uttered millions of times, but that does not remove them from the scope of human language; rather, it endows them with deeper meaning. Beyond a surplus of repetition, the contract corpora exhibit the properties of language; while certain edge cases may be amplified (exact sentence matches) and others reduced (hapax legomena), this is a difference in parameters, not a fundamental difference in form. For these reasons, the analysis of contract language remains intrinsically linguistic.

A more formal expression of the source of repetition in contract language is the fact that the speech act (Austin, 1962) performed by a contract itself is repetitive. While newswire and encyclopediae are focused on communicating new information to a reader, contracts are focused on creating arrangements between parties, similar to those created before. This is a very different illocutionary and perlocutionary act. Consequentially, contracts provide a new set of speech acts to study in NLP research, rarely seen in many of the genres of text frequently studied. However, a classification and examination of speech acts in both sets of corpora goes well beyond the scope of this study.

7 Conclusion

In this study, we documented the differences between contract documents and the sort ubiquitous in computational linguistics. Contract documents feature fewer hapax legomena, fewer pronouns, and much higher inter-sentence similarities; however, these similarities are not so redundant that the need for linguistic analysis is mitigated. This demonstrates both the need for new models of language specific to contract language and also the potential reciprocal benefits to research in linguistics and computational linguistics, as contract corpora can reduce the frequency of certain phenomena compared with the sort of corpora typically studied. We also hope in the future to potentially extend this analysis to other legal corpora and case reports.

Acknowledgments

Thanks to Tony Davis. Also, we would like to thank the reviewers, whose interest and feedback has inseparably made this work stronger.

References

Robert Anderson and Jeffrey Manns. 2017. Engineering greater efficiency in mergers and acquisitions. *The Business Lawyer*, 72:657–678.

Patrizia Anesa. 2007. Vagueness and precision in contracts: a close relationship. *Linguistica e filologia*, 24:7–38.

Shlomo Argamon, Paul Chase, and Jeff Dodick. 2005. The languages of science: A corpus-based study of

[8] A reviewer claimed that we were arguing against a strawman, and that no one would claim that contracts are not language. However, this is an actual claim—albeit one that was not subject to peer-review—made by a company in the legal technology space, "For the purpose of AI training, [technical legal] language cannot be considered a natural language. For contract review and approval, Natural Language Processing (NLP) and off-the-shelf solutions do not work." (`https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf`)

experimental and historical science articles. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27.

J.L. Austin. 1962. *How To Do Things With Words*. Harvard University Press.

Adelchi Azzalini and Antonella Capitanio. 1999. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

S. Bird, E. Loper, and E. Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Bjarne Blom and Anna Trosborg. 1992. An Analysis of Regulative Speech Acts in English Contracts - Qualitative and Quantitative Methods. *Quantitative and quantitative methods. Hermes (Aarhus)*, 82:83.

Daniel P. Broderick, Jonathan Herr, and Daniel E. Simonson. 2016. Method and System for Suggesting Revisions to an Electronic Document. U.S. Patent and Trademark Office, US20170039176A1/US10216715B2.

Luciana Carvalho. 2008. Translating contracts and agreements: a Corpus Linguistics perspective. *Avanĉos da linguística de Corpus no Brasil*, page 333.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2016. Information density and linguistic encoding (ideal). *KI - Künstliche Intelligenz*, 30(1):77–81.

Michael Curtotti and Eric C. McCreath. 2011. A Corpus of Australian Contract Language: Description, Profiling and Analysis. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 199–208, New York, NY, USA. ACM.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4):447–464.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 68–77, Vancouver, Canada. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Dorrit Faber and Karen Lauridsen. 1991. The compilation of a Danish-English-French corpus in contract law. *English computer corpora. Selected papers and research guide*, pages 235–43.

W. Nelson Francis and Henry Kučera. 1964, 1971, 1979. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown).* Brown University, Providence, Rhode Island, USA.

Michael AK Halliday. 1988. On the language of physical science. *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–178.

Zellig S Harris. 2002. The structure of science information. *Journal of biomedical informatics*, 35(4):215–221.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jason King. 2018. English Wikipedia Articles 2017-08-20 SQLite. Kaggle.

Julia Kristeva. 1980. *Desire in language: A semiotic approach to literature and art*. Columbia University Press.

Christian Mair. 1997. The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. *Trends in Linguistics Studies and Monographs*, 101:1537–1544.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Abdel Karim Mohammad, Nabil Alawi, and Maram Fakhouri. 2010. Translating contracts between english and arabic: Towards a more pragmatic outcome. *Jordan Journal of Modern Languages and Literature*.

Jane Norre Nielsen and Anne Wichmann. 1994. A frequency analysis of selected modal expressions in German and English legal texts. *HERMES-Journal of Language and Communication in Business*, 7(13):145–155.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.