A Corpus of Australian Contract Language

Description, Profiling and Analysis

Michael Curtotti*
School of Computer Science
Australian National University
Canberra, ACT, Australia
michael.curtotti@anu.edu.au

Eric C. McCreath†
School of Computer Science
Australian National University
Canberra, ACT, Australia
eric.mccreath@anu.edu.au

ABSTRACT

Written contracts are a fundamental framework for economic and cooperative transactions in society. Little work has been reported on the application of natural language processing or corpus linguistics to contracts. In this paper we report the design, profiling and initial analysis of a corpus of Australian contract language. This corpus enables a quantitative and qualitative characterisation of Australian contract language as an input to the development of contract drafting tools. Profiling of the corpus is consistent with its suitability for use in language engineering applications. We provide descriptive statistics for the corpus and show that document length and document vocabulary size approximate to log normal distributions. The corpus conforms to Zipf's law and comparative type to token ratios are consistent with lower term sparsity (an expectation for legal language). We highlight distinctive term usage in Australian contract language. Results derived from the corpus indicate a longer prepositional phrase depth in sentences in contract rules extracted from the corpus, as compared to other corpora.

1. INTRODUCTION

Contracts govern economic and cooperative transactions from trivial exchanges to major national infrastructure projects. Contract drafting and negotiation is thus a major vehicle of economic and societal activity. Any large organisation (whether private or public) must unavoidably invest significant resources in developing and concluding contracts - as the contracts it enters into define its legal relations with the organisations and individuals with which it interacts. As noted by Khoury and Yamouni, contracts are an integral part of any business enterprise and "it is difficult to overstate their importance to the business world"[27, p16].

Our ultimate purpose is to use the corpus to gain insight into the nature of contract language as an input to the development of software based drafting tools, particularly to assist drafters to identify and remove ambiguity in contracts.[1] Currently the tools available to most contract drafters consist primarily of Microsoft Word and perhaps libraries of contract templates. Drafters would benefit from software tools which specifically address their needs as contract drafters and negotiators. One example is a facility that detects and automatically highlights defined terms - to assist the drafter to properly use such defined terms. Well known forms of ambiguity such as prepositional phrase attachment ambiguity and conjunction ambiguity can easily enter contract text.

A contract corpus potentially also serves other purposes such as:

1. an empirical (particularly linguistic) exploration of contract language as a variety of English;
2. the automatic extraction of a domain ontology for contracts;
3. a differential comparison of Australian contract language with other forms of legal English (e.g. legislation) or contract language in other jurisdictions;
4. a quantitative assessment of whether actual contract language conforms to modern norms of "good" drafting practice as mandated by the plain English movement[49];
5. as an input for automatic contract management within organisations;
6. as an input for identification of contracts and the terms of contracts within the vast electronic document collections of large organisations; or
7. as an aid to translation of contracts from one language to another.[2]

*Michael Curtotti is undertaking a part-time Master of Philosophy through the School of Computer Science at the ANU. He works as a Senior Lawyer within the ANU Legal Office.

†Eric McCreath completed a PhD at the University of New South Wales in 1999 and is currently a Lecturer in the School of Computer Science at ANU.

ICAIL '11, June 6-10, 2011, Pittsburgh, Pennsylvania, USA

[1] Ambiguous drafting can result in loss and litigation for contracting parties. See for example `http://www.theglobeandmail.com/report-on-business/article838561.ece`, where the meaning of a provision with multimillion dollar implications for the parties turned on the placement of a comma.

[2] Examples of some of these applications can be seen in Section 2.

An initial use to which we have put the corpus is as a data source for machine learning for the purpose of multi-class classification of lines within contracts (enabling identification of entities such as headings, rules, definitions, parties and signature blocks)[11]. In this paper we report work on the description, profiling and initial analysis of a corpus of 256 Australian contacts. Initial analysis consists of chunking over a sub-corpus of rules extracted from the corpus.

In Section 2 we describe related work. Section 3 describes the design of the corpus. Section 4 outlines the tools and data used. Section 5 analyses the suitability of the corpus for its intended computational application. Section 6 reports chunking analysis to explore phrase occurrence. Section 7 provides conclusions and outlines potential future work.

2. RELATED WORK

Contracts are studied from a wide range of perspectives and disciplines. Most well known to the legal profession, the study of "contract law" is concerned with the laws or rules of contracting (including extensive legal thinking on the interpretation or "construction" of the meaning of contracts and the management of ambiguity in contracts in the context of legal disputes). Contracts have also been widely studied from the point of view of economic and social theory[39].

Corpus linguistics or natural language processing in relation to contracts falls within the broader application of such techiques to legal documents in general which has attracted extensive work. McCarty[33] for instance shows that state of the art statistical parsers can parse complex judicial pronouncements in a corpus of appelate judicial decisions. Moins and Boiy[36] apply classification to detect argument in text applying features such as n-grams, parts of speech tags and modal auxiliaries in a corpus including court decisions, parliamentary records and human rights advocacy web sites. Also related to such work is data mining or text mining in legal texts. Straneiri and Zeleznikow review the application of data mining techniques to legal documents, including techniques such as information extraction, text categorisation, text clustering and text summarisation[45, Chapter 8].

Application of such techniques to legislative documents (the texts of which more closely parallel the contractual domain) is also considerable. Bartolini[4], Francesconi[14], Mencia[34], Bacci et al.[3], Hasan et al.[22] and Biagioli et al.[5] carry out work in relation to classification of data within legislative texts. Venturi[48] undertakes work on the linguistic characterisation of legislative language for the purposes of computational semantic analysis. Van Gog and Van Engers[46] use natural language processing to convert legislative texts into 'objects' that can be represented using object modelling such as UML. Allen et al.[1] report on the use of "Aide" for the logical representation of legislative provisions. These references are indicative of the scope of such work, rather than comprehensive.

Research involving the specific application of computational techniques to contracts is more limited but in some cases substantial. Four fields of work are particularly noteworthy for the purposes of this paper:

1. the logical (or formal) representation of contracts rules;
2. the creation and implementation of e-contracts;
3. the linguistic study of contracts using corpora; and
4. natural language processing in application to contracts

Notable work has been carried out on the logical representation of contract rules [12, 17], and on the creation and implementation of electronic contracts[25]. Also work has been undertaken on developing XML representation for e-contract purposes.[3]

Little work, as far as we are able to determine, has been carried out on the natural language processing of contracts or in relation to the specific study of contract corpora. The following are the few examples of which we are aware. Blom and Trasborg[8] carry out an early study of a corpus of contracts, examining linguistic characteristics. Faber and Lauridsen[13] discuss the compilation of a corpus of contract law, a sub-component of which is a collection of contract texts. Their corpus has become known as the "Aarhus Corpus in Contract Law". Norre Nielsen and Wichmann[38] study the expression of 'obligation' in German and English in contract law corpora. Klinge[29] examines contractual modality from a pragmatic linguistics perspective. Anesa[2] studies vagueness and precision in contracts using a corpus of 12 contracts. Carvalho[10] studies a parallel corpus of English and Brazilian contracts with the purpose of increasing translation accuracy. Mohammad et al.[37] study a small parallel corpus of English and Arabic contracts again to improve translation accuracy. Indukuri and Krishna[25] carry out classification of clauses on a single contract. Varadarajan[47] reviews best practices in respect of text mining over business documents (including contracts). Minakov et al.[35] report contract template creation from the automatic clustering and semantic analysis of a collection of 25000 insurance documents in an insurance company. Sayeed et al.[42] develop a system for contract template complaince based on document similarity.

While each of the examples advance the study of contracts in particular areas, a coherent framework addressing the particular requirement for and character of the application of natural language processing or corpus linguistics to contracts does not emerge from the literature, rather one concludes that such study is very much in its early days and less developed for instance than the parallel work being undertaken in the legislative domain. Also, apart from the Aarhus corpus, which was compiled in the early 1990's, as far as we are aware, there is no publicly available corpus of contracts (or list for such a corpus) that could form the basis of study by a number of research groups. A likely reason for the slower development of this field is that until recently it would have been extremely difficult to obtain contract texts.

Also relevant to this paper is work on the design and profiling of corpora. Such work is referenced in context, in the sections which follow.

3. CORPUS DESIGN

The way in which a corpus is designed is heavily influenced by the purpose behind its creation: for example, whether it is being created as a general linguistic resource or to serve the needs of a specific project[44, p13][24, p26]. A general design principle to be derived from such a statement therefore is that a corpus should be designed to be suitable for its intended purpose.

Given our ultimate research aim is deployment of software tools operating on individual contract drafts, the selection

[3]http://docs.oasis-open.org/legalxml-econtracts/CS01/legalxml-econtracts-specification-1.0.pdf

of material to comprise the corpus is straightforward: i.e. complete written contracts. This selection of texts meets a fundamental requirement of a corpus: i.e. that it represent a language or some part of a language[6, p246]: in this case contract language.

We further limit the corpus to *Australian* contract texts. As different jurisdictions have different laws, this can be expected to influence the character of contract language used in that jurisdiction. Also different English speaking jurisdictions (e.g. U.S. versus Australia) have developed significantly different contracting styles. Distinguishing between different jurisdictions will enable future studies carrying out empirical comparisons of these jurisdictional differences.[4] Also, limiting the corpus to one jurisdiction removes such differences which can be expected to complicate the development of a representative corpus.

In order to compile our corpus a search was undertaken on the Google Australia webpage [5] using the search terms: 'clause party agreement'[6], with the search limited to 'pages from Australia' and the filetype limited to '.doc'. Using the selected generic search terms minimizes biasing to any particular contract types (for instance employment contracts or intellectual property contracts). The limitation to '.doc' files, flows from Microsoft Word being the primary tool used within the legal industry for document creation,[7] and the intended deployment of software tools within that context. Each document was visually inspected by one of the authors to verify that it constituted an example of an Australian contract and documents were added to the corpus in order of their appearance in the Google search results until the corpus was approximately 1,000,000 words in size. This resulted in a corpus of 256 contracts. The collection of the corpus was undertaken in the period 6 - 24 December 2009 and a listing of the urls is made available over the web, to facilitate similar research.[8]

One shortcoming of compiling a corpus from publicly available sources on the web is that it will not capture contracts that owners consider to be sensitive and therefore do not make public. Further many of the contracts included in the corpus are in the form of contract templates and are in minor respects not complete (e.g. containing fields that need to be completed when the contract is deployed in practice). This is not necessarily a disadvantage in developing a tool for contract drafters, as such constructs and drafts in various stages of completion would need to be dealt with by a drafting tool. Nonetheless, many of the included examples have not undergone a process of negotiation to a concluded agreement. The language represented in our corpus is thus more typically that appearing in contract templates rather than executed contracts. Further we may assume that public organisations will be more ready to publish copies of their legal instruments rather than private organisations. This is borne out, for instance, by the high occurrence of terms such as 'university', in the corpus. While such factors need to be borne in mind in basing conclusions on the corpus, these considerations are not significant in the context of our project aims: particularly in a context where very little is available in the way of accessible corpora of contracts.

4. TOOLS AND MATERIALS

In order to carry out the analysis reported here, we used the Natural Language Toolkit (NLTK)[7] (which provides a wide variety of highly accessible tools and corpora for natural language processing) and MontyLingua[31] (an end to end parts of speech tagging and chunking tool). Python was used to develop a number of corpus related utilities to assist in the extraction of data and calculation of results.[9]

To undertake comparative analysis of the contract corpus, we used a number of corpora available through NLTK: the Brown corpus (intended to be a representative sampling of written American English and composed of 500 tracts of around 2200 words)[15]; the Reuters corpus composed of Reuters news wire reports; a corpus of ABC science and rural news articles; Jane Austen's Emma extracted from `http://www.gutenberg.org`; and a corpus of movie reviews.

We used the Weka Data Mining Software to carry out classification of rules from non-rules[21].

5. PROFILING: SUITABILITY FOR LANGUAGE ENGINEERING

A central question in the use of corpora for language engineering is whether the corpus in question is representative of the population from which it is drawn[32, p119]. As the population is often extremely large (in this case the population of all Australian contract texts), directly answering this question is difficult. We follow Sarkar and others in applying an indirect method of 'fast profiling' a corpus to assess its suitability for language engineering[41][18]. This method (as we have applied it) consists of the following stages:

1. developing a 'rough profile' of the corpus reporting key statistical and numeric measures;
2. manual sampling to check for obvious idiosyncracies; and
3. the application of diagnostic tests for sparseness such as non-conformance with Zipf's law and low type-to-token ratio.[10]

Manual sampling is addressed below in the context of an examination of token and collocation frequencies, focussing

[4]Based on the authors' domain knowledge U.S contracting styles, for instance, appear significantly different to Australian styles in respect of a range of features including sentence lengths, formality of lexicon and use of subparagraphing.

[5]`http://www.google.com.au`

[6]By a process of trial and error we found that this particular search combination returns research results with a higher density of contract documents in the search results.

[7]See for instance surveys undertaken by the International Legal Technology Association report[16] that 96% of law firms use various versions of Microsoft word as their primary word processing software. Given the prevalence of this format focussing on it enables future software development to take advantage of information embedded in the format.

[8]The list is available at: `http://cs.anu.edu.au/~Michael.Curtotti/`.

[9]At `http://cs.anu.edu.au/~Michael.Curtotti/` we make available three python files used in research related to this paper: a set of corpus utilities, a rule based line tagger for characterising lines in contracts, and a feature extractor used for machine learning. We also make available the dataset used in work associated with our classification paper[11].

[10]Sarkar et al. apply also a fourth set of tests related to the use of function words, which we do not reproduce here.

on terms identified as most characteristic of the contract corpus. (See Sub-section 5.1 and following.)

5.0.1 Descriptive Statistics

Table 1 provides basic statistical measures for the contract corpus. The corpus is constituted of approximately 1,000,000 words (the same scale as the Brown and Reuters corpora).[11]

Table 1: Basic Statistical Measures

Corpus Properties	Value
No of documents	256
Corpus length in tokens	1043364
No of distinct tokens	14217
Av. document length	4075.64
St. dev of doc length	3629.76
Skew of doc length	2.89
Av. no of distinct tokens per doc	704.40
St. dev of distinct tokens per doc	345.88
Skew of distinct tokens per doc	1.60

The measures reported above go beyond the Sarkar et al. methodology, as we also examined skew in document length. Our sample showed a significant right skew. This is explained as a lognormal distribution, which is characteristic of a number of linguistic features. Document length in a corpus, for example, can be approximated by a lognormal distribution. Word length and sentence length are also lognormally distributed[43].[12] In general, skewed distributions are particularly common where the average of a data set is low, variance of individual data points high and values cannot be negative[30]. The skew in contract document length is consistent with our intuitions about contracts and suggest that an unbiased sampling of contracts would have such a characteristic. Contracts (typically) are not long (anecdotally being say 2 to 10 pages in length), although larger (rarer) projects or complex relationships may be accompanied by significantly longer documents, sometimes running to many dozens of pages. In the contract corpus, document length and vocabulary length conform approximately to a lognormal distribution (See Figures 1 and 2).

The value of considering the nature of the probability distribution the data exhibits is illustrated by noting that given document length is approximately lognormally distributed we are able to apply the geometric mean (3125) and the standard deviation of the log transformed values to derive a figure for a 68% confidence interval of document length (between 1543 and 6236 tokens) and 95.5% confidence interval (between 762 and 12808 tokens).[13] We may conclude that the length of Australian contracts (if our sample is representative) are highly likely to be in this order i.e. between 700 and 13000 words in length: a result relevant to the computational performance that we may encounter in carrying out many NLP related tasks.

For the purposes of assessing the suitability of the contract corpus for language engineering, these descriptive statistics do not suggest any issue in the sampling of the corpus.

Table 2: Lognormal and Related Measures

property	doc length	doc vocab
Geometric mean	3125.21	633.29
Median	2916.50	622.50
Log mean	3.49	2.80
Log st. deviation	0.31	0.20
Log Skew	0.27	-0.00412

5.0.2 Type to Token

An examination of the type to token ratio of the contract corpus establishes that the corpus is significantly less sparse than either the Brown or Reuters corpora, implying a reduction in sparsity issues as compared to those corpora. Table 3 shows type to token ratios for different sizes of sub-corpora drawn from these three sources, from 100 to 1000000 tokens. The comparison moreover is consistent with what we would expect: that the vocabulary of contracts would be less diverse than that of news articles, which would be less diverse than that of general English. Column 4 in Table 3 reproduces figures for type to token ratios derived by Sarkar et al.[41], which although of the same order of magnitude are not identical. The comparison is provided with the qualification that given the use of different software and processing methods, some difference in results is to be expected.

Table 3: Type to Token Ratios

Length	Contract	Reuters	Brown	Brown[41]
100	1.72	1.47	1.56	1.449
1600	4.19	2.65	2.57	2.576
6400	6.11	4.05	3.60	4.702
16000	9.03	5.69	4.69	5.928
20000	9.39	6.17	4.98	6.341
200000	30.07	18.45	9.89	n/a
1000000	71.74	41.05	21.64	20.408

5.0.3 Zipf Curve

Each word in a corpus has a particular frequency. Zipf's law (which Zipf applied to a wide variety of phenoma) in

[11] In extracting these measures all tokens were used (i.e. no filtering was applied to remove punctuation tokens or stop words). The only preprocessing applied to measure the vocabulary size, was conversion of all terms to lower case. Stemming was not applied.

[12] Interestingly the lognormal distribution, despite its relevance to linguistic phenomena, barely finds mention in relevant articles and does not appear at all in Manning[32] or Jurafsky[26] (both standard texts in computational linguistics). An interesting instance in the legislative field where we do find the lognormal mentioned is in the work of Bommarito and Katz [9], who examine the properties of the citation network within the United States legal code (i.e. cross references from one section to another), finding that the distribution of the number of cross-references from one section to another (normalised for section length) follows a log normal form.

[13] For example the 95.5% confidence interval can be obtained by adding two times the log standard deviation to the log mean for the upper bound and subtracting the same amount for the lower bound. The resulting figures are converted back to counts by exponiation. (See explanation in Limpert et al.[30]).

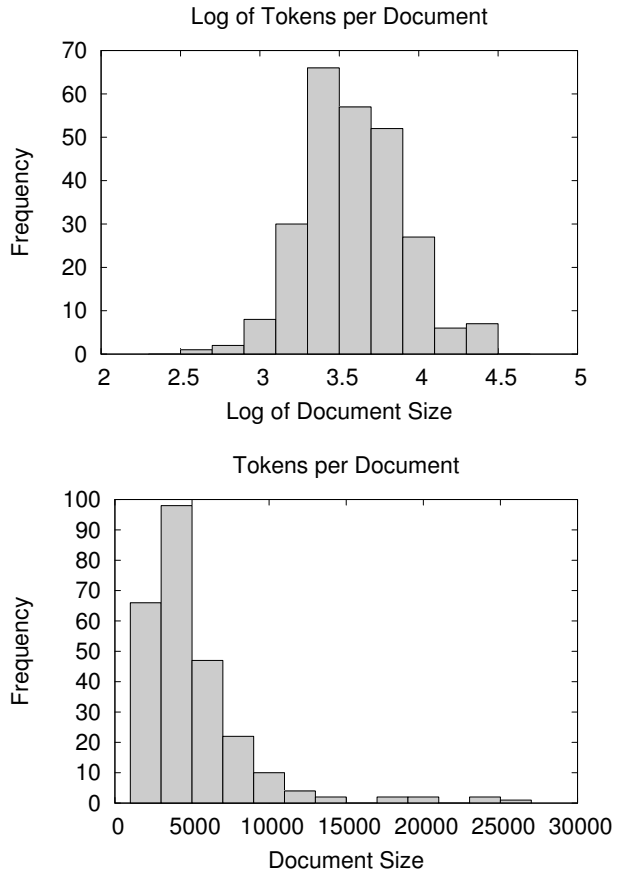

Figure 1: Histograms showing lognormal distribution of document length

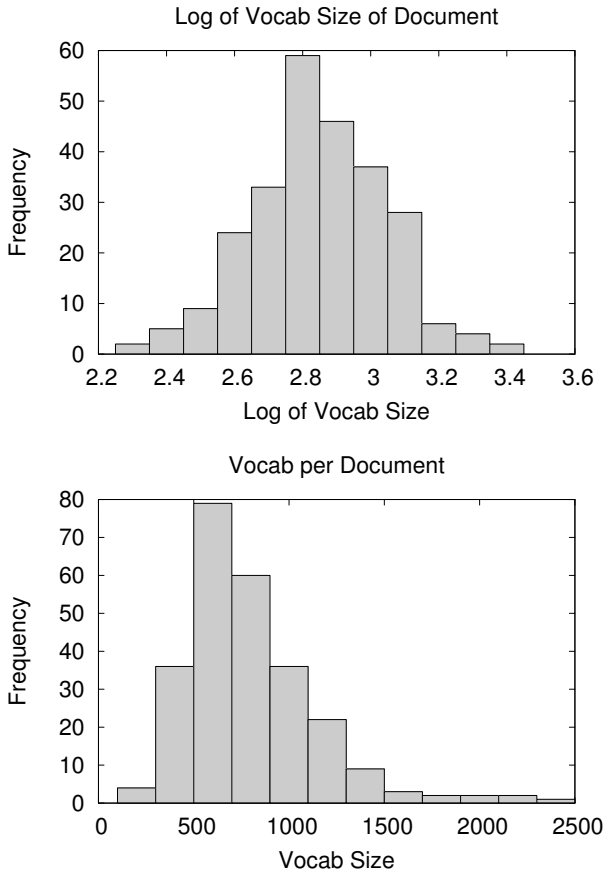

Figure 2: Histograms showing lognormal distribution of document vocabulary size

respect of lanuage holds that frequency of a term in a corpus is inversely proportional to its rank order[32, pp23-25]. Failure to conform to this law may indicate that the sample is unrepresentative.

A Zipf chart for a corpus that conforms to Zipf's law (comparing log of rank to log of frequency) should roughly approximate a line with a slope of -1[18], although Ha et al.[19] examining larger corpora finds that the slope for languages such as English and Spanish drop to about -2 for rank above 5000 (a result which also seems to hold for the contract corpus). In related work Ha et al.[20], combining frequencies of n-grams as 'units of meaning' in languages such as English and Chinese, show that Zipf law for English is maintained at a slope of -1 if n-grams larger than one are accounted for. Note that the Zipf curve for the Brown corpus shows the same characteristic as reported here for the contract corpus (i.e. deviation to a steeper slope above a certain rank (5000 in that case)[20]. The contract corpus thus comfortably conforms to Zipf's law, as illustrated in Figure 3.

5.1 Token Occurrence

Information about the most frequent terms in a corpus does not necessarily identify the terms that best characterise the corpus, as compared with other language usage. Deriving a comparative measure provides information as to what makes a corpus distinctive: in this case what is distinctive about contracts. Such a list of 'distinctive terms' also enables an easy visual inspection of whether high 'distinctive terms' are out of place. A number of measures might be applied to this task including Pearson's chi squared ratio, Mann-Whitney's frequency ranking and log-likelihood ratio ('the goodness-of-fit') test[28].

Rayson and Garside[40] employ the log likelihood ratio on the basis that it does not assume a normal distribution and does not have the same difficulties as the chi-squared test in respect of low frequency values. Applied to words, the method calculates the log likelihood ('LL') ratio of the frequency of a word in frequency lists extracted from each corpus. The method results in a ranking of words according to their LL ratio, thus highlighting the most significant term differences between the corpora. Such differences when comparing a specialised language to general English, may assist us in identifying special features of the corpus that may impact on language engineering.[14]

In applying this method here, first, the 500 most frequent

[14]LL is calculated using the formula: $LL = 2(a \log(\frac{a}{E1}) + b \log(\frac{b}{E2}))$ where $E1 = c(\frac{a+b}{c+d})$ and $E2 = d(\frac{a+b}{c+d})$ and a and b are the frequency of the subject word in the corpora being compared and c and d are the total number of tokens in the corpora being compared.

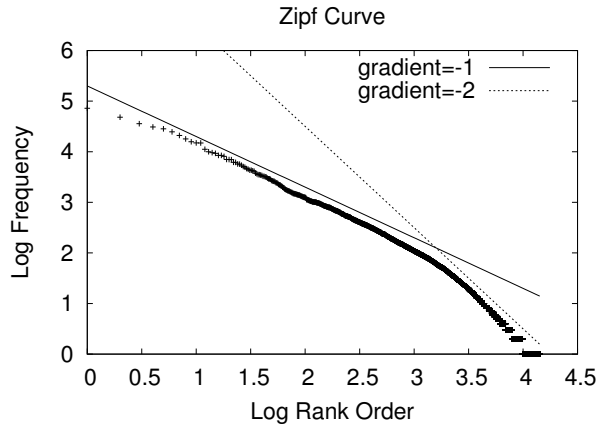

Figure 3: Zipf Curve for Contract Corpus. The curve maintains a slope of -1 until exceeding rank 1000 when it begins to deviate to an apparent slope of -2.

terms were extracted from the contract corpus. This limits the sample to terms which occur with some frequency in contracts: with the least most common term in the list occurring 249 times in the contract corpus. Looked at another way, this list captures approximately 816000 of the terms used, or 78%, of the term usage occurring in the corpus.

LL measures were then derived in comparison with the Brown and Reuters corpora. Table 4 shows the highest ranked terms (after removal of punctuation) for LL in its first column. Rayson and Garside describe log likelihood in the following terms:

> "[Log Likelihood has] the effect of placing the largest LL value at the top of the list representing the word which has the most significant relative frequency difference between the two corpora ... words which appears with roughly similar relative frequences in the two corpora appear lower down the list." [40]

In mathematical terms the measure provides similar results to taking the absolute value of the difference between the frequencies in the two corpora (as shown in column 2 of Table 4).

Manning illustrates a slightly different measure (the ratio of the frequency of a given term in two corpora i.e. frequency 1 / frequency 2) "since they can be interpreted as likelihood ratios" [32, p 175]. Column 3 shows the highest ranked terms produced utilising this measure. Notably the terms identified in this case are quite different. Visual inspection suggests that this simpler metric is rich in terms of the subject matter of the corpus with the terms identified being such as might far more readily lead one to conclude the list comes from a set of legal documents. It might be a good measure for instance for ontology extraction or for identifying distinctive document vocabulary.

A preprocessing step that is sometimes applied when using log likelihood is the removal of material such as 'function words' by using a 'stop list' (For example see He et al.[23]). Such a preprocessing step does not appear to be relevant when taking a simple ratio of frequencies.

Table 4: Most Distinctive Terms.

CtoB Log L.	abs(C - B)	C/B
(+) **or**	**or**	organisation
(+) **agreement**	**any**	gst
(-) **was**	the	authorised
(+) **any**	**agreement**	licence
(-) **his**	**was**	provider
(+) **party**	**his**	software
(+) **clause**	(-) a	mediation
(+) **shall**	**it**	invoice
(+) parties	(+) by	mediator
(-) **it**	(+) this	copyright
(+) **information**	to	licensee
(-) **but**	**party**	waiver
(+) date	(+) will	abn
(+) services	**shall**	dva
(-) they	**but**	funding
(+) under	**clause**	ip
(+) schedule	(+) of	licensor
(+) project	**information**	nrl
(-) would	**under**	**clause**
(+) commonwealth	(+) other	confidentiality

A "(+)" indicates a higher occurrence in the contract corpus while a "(-)" indicates a lower occurrence. Bolding highlights terms which co-occur in the first column and the second or third column.

The first and second columns are also informative however. For instance the word 'or' appears far more frequently in the contract corpus than the Reuters or Brown corpora: i.e. a frequency of 20.077 to 1.887 to 3.622. The determiner 'any' also appears far more frequently in the contract corpus. By contrast the past tense 'was', the pronoun 'his' (in relation to the Brown corpus) and the pronoun 'it', all appear less frequently. Although not shown in Table 4, commas also have a different usage in contracts being used about half as frequently as in the Reuters or Brown corpora, while colons occur around five times as frequently.

Each of these observations suggest how such a list may be used for further investigation of the contract corpus - with frequency difference serving as a marker for differences in language usage that may potentially be significant to the intended language engineering application: e.g. investigating differences in disjunction, the use of tense or the use of pronouns. In an experiment to classifying lines as 'rules' or 'non-rules' using 1-grams as learning features, we found terms such as 'the', 'any', 'and', 'to', 'may', 'that', 'or', 'must' and 'will' to be key features for the classification (with such terms marking the occurrence of rules).[15] A number of these terms are also distinctive of contracts as a whole.[16]

Using domain knowledge we may also look for frequency differences in what we may intuitively consider to be 'key terms' in contracts. A short list of such terms might include the words 'if', 'means', 'must', 'may' and 'where'. The word 'means' is a marker for definitions[17], while the words 'must'

[15]This experiment was carried out using the weka data mining software[21].

[16]Note that as the terms clause, agreement and parties were used for document selection they are their frequency is discounted as informative as their frequency is determined by the sampling method.

[17]Apart from domain knowledge that would suggest this, in experiments we have carried out using n-grams as features

and 'may' are used respectively as markers for obligation and freedom. The words 'if' and 'where' are used to mark conditionality in contracts. Table 5 illustrates the higher frequency of these terms in the contract corpus as against either the Brown or Reuters Corpora (columns 3-5). Columns 6 and 7 show that taking a simple difference in frequencies, as compared to log likelihood gives a notably higher ranking to these terms.

Table 5: Key Term Measures.

Contract frequency rank	Term	Contract frequency (per 1000 tokens)	Reuters frequency	Brown frequency	Contract to Brown Log Likelihood	Contract to Brown abs freq. diff.
37	**if**	3.4	0.9	1.9	210	61
59	**may**	3.5	1.2	1.2	71	43
36	**must**	2.3	0.2	0.9	137	65
55	**means**	2.1	0.1	0.3	48	52
85	**where**	1.4	0.2	0.8	380	197

5.2 Collocations

Collocations found in the contract corpus (extracted using NLTK) are found to contain common legal terms of art or contractual phrases. Terms such as: intellectual property; confidential information; third party; written consent; tax invoice; written notice; without limitation; property rights; dispute resolution; force majeure; personal information; business day; taxable supply; good faith; moral rights; and governing law all appear among the 50 most frequent collocations. The same list however also contains some collocations which are clearly specific to particular documents e.g. nemde solver; flight attendant; mobile phone; nrl club and rugby league.

5.3 Profiling Results

The foregoing 'profiling' of the corpus establishes its validity of its design for the purposes of language engineering. Moreover in carrying out this profiling aspects of the exploration which were of interest to us emerged. These we have noted in the discussion above: the log-normal distribution of length and vocabulary of documents in a corpus (which is found to hold in respect of the corpus), the deviation of the Zipf curve for lower ranked terms (a pattern seen to hold for English corpora generally but resolved if n-grams higher than one are taken into account). We considered what measures might prove most useful in indentifying distinctive term occurrence - noting differences in various mathematical measures of distinctiveness. Terms identified as distinctive of contracts included both function terms and terms that domain knowledge might suggest would be distinctive.

for classification we have found the word 'means' to be the most the most effective n-gram feature when seeking to classify lines containing definitions as opposed to other text in contracts. This experiment was carried out using the weka machine learning software[21].

6. CHUNK ANALYSIS

We also undertook chunk analysis to explore phrase occurrence in the contract corpus particularly in comparison to related work by Venturi[48] who carries out a study of Italian and English legislative language as against general language. Her key finding is a higher occurrence of prepositional phrases and finite verb phrases in both Italian and English legislative texts. The question we explored was whether similar phrase occurrence patterns apply in respect of our corpus of contracts. Venturi's study was carried out using a chunking approach, which we also adopted.

As a first step a sub-corpus of 50 contracts constituted of 'contract rules' was extracted and hand tagged to classify the content according to whether it constituted substantive legal content (i.e. clauses and definitions) or 'non-rule' material (such as headings, tables of contents, execution blocks, etc). All non-rule material was stripped from this sub-corpus.

MontyLingua was used to apply parts of speech tags and to chunk the sub-corpus. Comparison was then undertaken between this sub-corpus and six other corpora (all available through NLTK): the Brown, Reuters, ABC (divided by rural and science reports), Emma by Jane Austen and Movie Reviews. Table 6 shows results for all corpora. The first seven rows show occurrence per thousand tokens. The bottom 7 rows show occurrences per sentence. For all corpora, except the Brown corpus, the occurrence of prepositional phrases was notably higher in the contract corpus than other corpora. For instance as compared with general or popular language (Jane Austen and movie reviews) prepositional phrase occurrence was 55.6% higher. As against news corpora the occurrence was also higher (though only around 25%). The Reuters and Brown corpora show around the same occurrence of verb phrases, other corpora having a higher occurrence of verb phrases (both finite and infinitive). These results (in respect of prepositional phrases) are in the same direction as the findings reported by Venturi (for instance she finds a 36% higher occurrence of prepositional phrases in a corpus of environmental law as opposed to the Wall Street Journal).

Sentences in the contract corpus are longer than in the other corpora and as a consequence there are more prepositional phrases per sentence.

Venturi also studies the prepositional phrase chain depth of legislative versus general language finding a greater depth in legislative language. Figure 4 shows similar results to those found by Venturi: i.e. prepositional phrase length is not only longer on average, the proportion of sentences having a higher prepositional phrase depth is higher for contract language in our sub-corpus as compared to general language. The only corpus which approached the contract sub-corpus, was the writings of Jane Austen (notably a somewhat older corpora). The contract corpus has sentences of very high length. A visual inspection of such sentences shows them essentially to be long lists (e.g. lists of definitions separated by semi-colons or lists of conditional rules separated by semi-colons).

7. CONCLUSIONS

We have reported the design and profiling and phrase

Table 6: Chunk Occurrence.

	C	B	R	A-S	A-R	JA	MR
NP	231	721	220	233	233	219	220
PP	126	117	91	103	98	81	81
VP	99	92	88	118	118	126	115
Adj	15	13	10	18	14	37	28
FV	87	81	77	101	101	102	98
IV	12.1	11	11.3	17.5	17.2	23.8	17.2
S	22.9	27	32.0	36.3	36.3	30.5	47.9
tok/s	43.5	37	31.2	27.5	27.6	32.8	20.9
PP/s	5.5	4	2.8	2.8	2.7	2.7	1.7
NP/s	10	9	6.9	6.4	6.4	7.2	4.6
VP/s	4.3	3	2.8	3.3	3.3	4.1	2.4
FV/s	3.8	3	2.4	2.8	2.8	3.4	2.0
IV/s	0.5	0	0.4	0.5	0.5	0.8	0.4
A/s	0.6	1	0.3	0.5	0.4	1.2	0.6

Code: NP = Noun Phrases per thousand words; PP = Prepositional Phrases per thousand words; VP = Adjectival Phrases per thousand words; FV = Finite Verb Phrases per thousand words; IV = Infinitive Verb Phrases per thousand words; S = average sentence length; tok/s = tokens per sentence; etc.

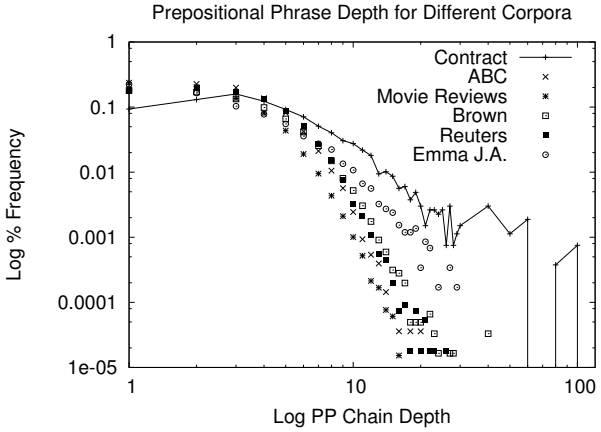

Figure 4: Log of % Frequency of Prepositional Phrase Depth for Different Corpora

analysis of a corpus of Australian contract language, including comparisons with other corpora. Profiling supports the validity of the method employed in compiling the corpus from the web and highlights interesting results in respect of it: e.g. conformance with Zipf's law, a lognormal distribution for document length and vocabulary. The corpus has lower sparsity than reference corpora such as Brown and Reuters.

Initial work is reported in the identification of distinctive contract terms at word and collocation level. A number of measures are explored for identifying such language.

Chunk analysis of the contract corpus highlights a number of features relevant to language engineering which echo findings of Venturi in relation to legislation: contract language displays a higher use of prepositional phrases, longer prepositional chain depth per sentence, and lower relative usage of verbs at a sentence level.

The work reported in this paper contributes to an end objective of developing NLP based methods to deliver contract drafting tools. It also provides an initial study of Australian contract language, and reports methods and sources that may be used for further corpus based studies by the authors or others.

In the next stage of work in relation to the corpus we plan to examine the use of defined terms in contracts and explore issues such as their formal representation and ambiguity detection in definitions.

8. REFERENCES

[1] R. Allen, P. Chuns, A. Mowbray, and G. Greenleaf. AustLll's aide — natural language legislative rulebases. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 223–224. ACM, 2001.

[2] P. Anesa. Vagueness and precision in contracts: a close relationship. 2007.

[3] L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54, 2009.

[4] R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, and C. Soria. Automatic classification and analysis of provisions in italian legal texts: a case study. In *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, pages 593–604. Springer, 2004.

[5] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 133–140. ACM, 2005.

[6] D. Biber, S. Conrad, and R. Reppen. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, 1998.

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

[8] B. Blom and A. Trosborg. An analysis of regulative speech acts in English contracts-Qualitative and quantitative methods. *Hermes*, 9:83–112, 1992.

[9] I. Bommarito, J. Michael, and D. Katz. Properties of the United States Code Citation Network. *Arxiv preprint arXiv:0911.1751*, 2009.

[10] L. Carvalho. Translating contracts and agreements: a Corpus Linguistics perspective. *Avanços da linguística de Corpus no Brasil*, page 333, 2008.

[11] M. Curtotti and E. McCreath. Corpus Based Classification of Text in Australian Contracts. In *Australasian Language Technology Association Workshop 2010*, page 18.

[12] A. Daskalopulu and M. Sergot. The representation of legal contracts. *AI & Society*, 11(1):6–17, 1997.

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

[14] E. Francesconi. The Norme in Rete Project: Standards and Tools for Italian Legislation.

International Journal Legal Information, 34:358, 2006.

[15] W. N. Francis and H. Kucera. A Standard Corpus of Present-Day Edited American. Revised 1971, Revised and Amplified 1979. Department of Linguistics, Brown University Providence, Rhode Island, USA. `www.hit.uib.no/icame/brown/bcm.html`, 1964.

[16] C. Gibney and T. Corham. International legal technology association (ILTA) 2008 technology survey. `www.iltanet.org/WhitePaperPDFs/ 2008TechnologySurvey.aspx`, August 2008.

[17] G. Governatori. Representing business contracts in RuleML. *International Journal of Cooperative Information Systems*, 14(2-3):181–216, 2005.

[18] A. Goweder and A. De Roeck. Assessment of a significant Arabic corpus. In *Arabic NLP Workshop at ACL/EACL*, 2001.

[19] L. Ha, D. Stewart, P. Hanna, and F. Smith. Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics*, 1(8):1–12, 2006.

[20] L. Q. Ha, E. Sicilia-Garcia, J. Ming, and F. Smith. Extension of Zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–6. Association for Computational Linguistics, 2002.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software. *SIGKDD Explorations*, 11(1), 2009.

[22] I. Hasan, J. Parapar, and R. Blanco. Segmentation of legislative documents using a domain-specific lexicon. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, pages 665–669, 2008.

[23] T. He, X. Zhang, and X. Ye. An Approach to Automatically Constructing Domain Ontology. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 2006.

[24] S. Hunston. *Corpora in Applied Linguistics*. Cambridge University Press, 2002.

[25] K. V. Indukuri and P. R. Krishna. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference*, pages 1–5, New York, NY, USA, 2010.

[26] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edition, 2009.

[27] D. Khoury and Y. S. Yamouni. *Understanding Contract Law*. Lexis Nexis Butterworths, 7th edition, 2007.

[28] A. Kilgarriff. Comparing Corpora. *International Journal of Corpus Linguistics*, 6:1:97–133, 2001.

[29] A. Klinge. On the linguistic interpretation of contractual modalities. *Journal of Pragmatics*, 23(6):649–675, 1995.

[30] E. Limpert, W. A. Stahel, and M. ABTT. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51(5), 2001.

[31] H. Liu. Montylingua: An end-to-end natural language processor with common sense. `web.media.mit.edu/ hugo/montylingua`, 2004.

[32] C. D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 2000.

[33] L. McCarty. Deep semantic interpretations of legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 217–224. ACM, 2007.

[34] E. L. Mencía. Segmentation of legal documents. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 88–97. ACM, 2009.

[35] I. Minakov, G. Rzevski, P. Skobelev, and S. Volman. Creating contract templates for car insurance using multi-agent based text understanding and clustering. *Holonic and Multi-Agent Systems for Manufacturing*, pages 361–370, 2007.

[36] M. Moens, E. Boiy, R. Palau, and C. Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM, 2007.

[37] A. K. Mohammad, N. Alawa, and M. Fakouri. Translating contracts between English and Arabic and Arabic: Towards a more pragmatic outcome. *Jordan Journal of Modern Languages and Literature*, 2:1–28, 2010.

[38] J. Norre Nielson and A. Wichmann. A frequency analysis of selected modal expressions in German and English legal texts. *Hermes*, 13:145–155, 1994.

[39] J. Paterson, A. Robertson, and P. Heffy. *Principles of Contract Law*. Lawbook Co., 2nd edition, 2005.

[40] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora-Volume 9*, pages 1–6. Association for Computational Linguistics, 2000.

[41] A. Sarkar, A. De Roeck, and P. Garthwaite. Easy Measures for Evaluating non-English Corpora for Language Engineering: Some Lessons from Arabic and Bengali. Technical report, Technical Report 2004/05, Open University–Department of Computing, 16th February, 2004, 2004.

[42] A. Sayeed, S. Sarkar, Y. Deng, R. Hosn, R. Mahindru, and N. Rajamani. Characteristics of document similarity measures for compliance analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1207–1216. ACM, 2009.

[43] M. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PloS one*, 4(4):5372, 2009.

[44] J. Sinclair. *Corpus Concordance Collocation*. Oxford University Press, 1991.

[45] A. Stranieri and J. Zeleznikow. *Knowledge discovery from legal databases*. Kluwer Academic Pub, 2005.

[46] R. Van Gog and T. Van Engers. Modeling legislation using natural language processing. In *IEEE International Conference of Systems Man and Cybernetics*, volume 1, pages 561–566, 2001.

[47] S. Varadarajan, K. Kasravi, and R. Feldman. Text-Mining: Application Development Challenges. In *Applications and innovations in intelligent systems X:*

proceedings of ES2002, the twenty-second SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence, page 247. Springer Verlag, 2003.

[48] G. Venturi. Parsing legal texts. A contrastive study with a view to Knowledge Management Applications. In *Language Resources and Evaluation LREC 2008 Workshop on the Semantic Processing of Legal Texts*, page 1, 2008.

[49] C. Williams. Legal English and Plain Language: An Introduction. *ESP Across Cultures*, 1:111–124, 2004.