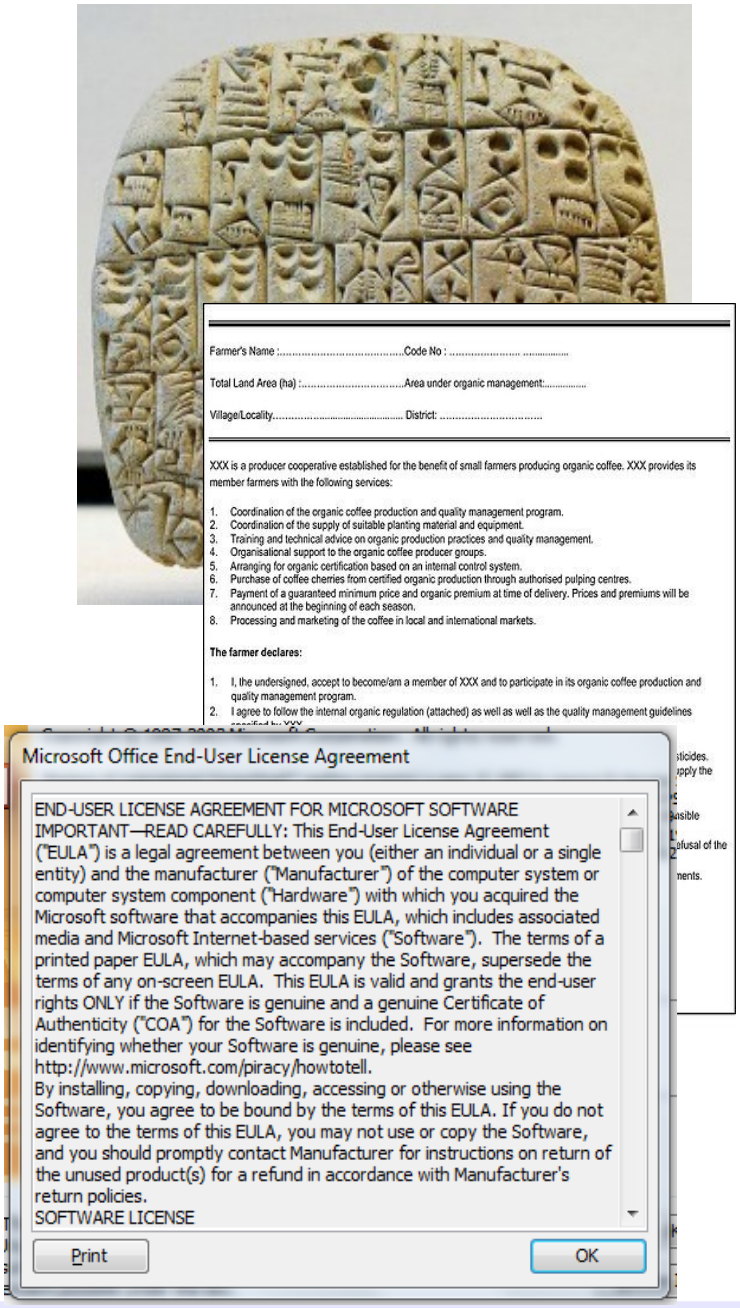
A Corpus of Australian Contract Language

Description, Profiling and Analysis

Michael Curtotti, Eric McCreath
Research School of Computer Science
Australian National University
Canberra, ACT, Australia

michael.curtotti@anu.edu.au,
eric.mccreath@anu.edu.au

Contracts



- There are many words that convey nuances of the idea of a contract:
  agreements, arrangements, accords, treaties, pacts, partnerships, marriages, alliances, wills, deals, oaths, threats, settlements, ultimatums, terms, conditions, laws, statutes, bargains, guarantees, awards, warranties, promises, pledges, undertakings, vows, assurances, engagements, requirements, demands, truces, cease-fires, compromises, mortgages, indentures, etc..
- Contracts are a fundamental to organisational and individual relationships and transactions.
- Contract drafting is a major economic activity for the legal industry.
- Frederick Sawyer:
  "Contract: an agreement that is binding on the weaker party."
  http://en.wikipedia.org/wiki/Contract :
  "A contract is a legally enforceable agreement between two or more parties with mutual obligations."

Corpus Design

Australian National University

- The design of a corpus is heavily influenced by the purpose behind its creation.
- We want to gain an insight into the nature of contract language as an input to the development of software based drafting tools.
- The corpus gives us a base for exploring a number of possibilities including:
  - segmenting contracts,
  - automatically highlighting defined terms,
  - finding dependency in definitions and providing a visualization of the dependency graph, and
  - assisting drafters identifying and removing ambiguity.

Corpus Design

Australian National University

A contract corpus potentially also serves other purposes such as:

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

Corpus Design

Australian National University

- Limited to Australian contracts.
- Google search:
  - phrase “clause party agreement”,
  - Pages from Australia,
  - “.doc” files only - 96% of lawyers use M$ word.
- Documents visually inspected.
- Collected until we reached 1,000,000 words, giving 256 documents.
- URL list publicly available:

  http://cs.anu.edu.au/~Michael.Curtotti
- In the process of attempting to make the raw documents available in some form.

Basic Statistical Measures

Australian National University

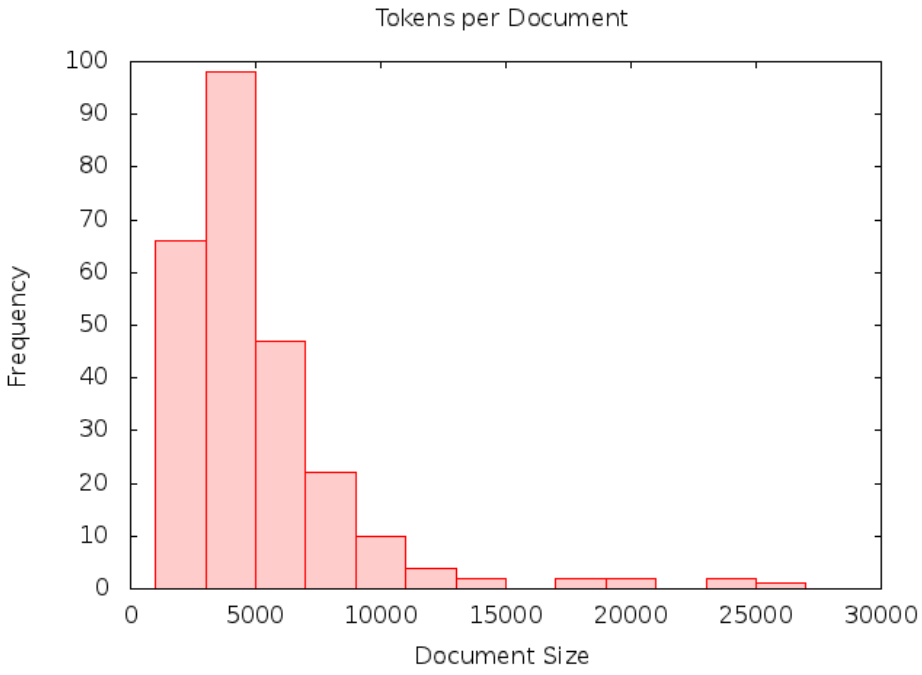
Corpus Properties	Value
No of documents	256
Corpus length in tokens	1043364
No of distinct tokens	14217
Av. document length	4075.64
St. dev of doc length	3629.76
Skew of doc length	2.89
Av. no of distinct tokens per doc	704.40
St. dev of distinct tokens per doc	345.88
Skew of distinct tokens per doc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

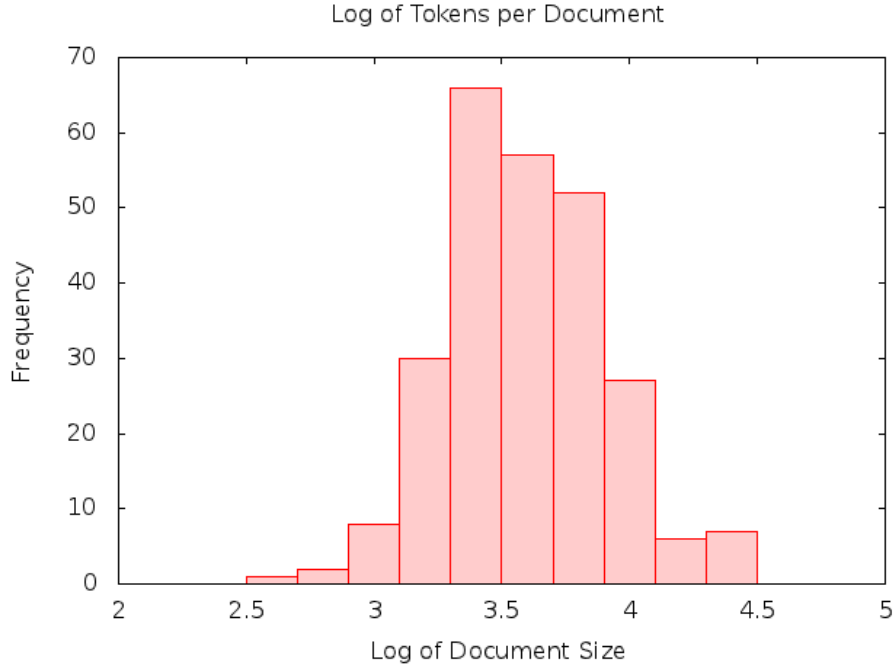
Fast Profiling

Australian National University

- We follow Sarkar and others in applying an indirect method of 'fast profiling' a corpus to assess its suitability for language engineering.

Tokens per Document





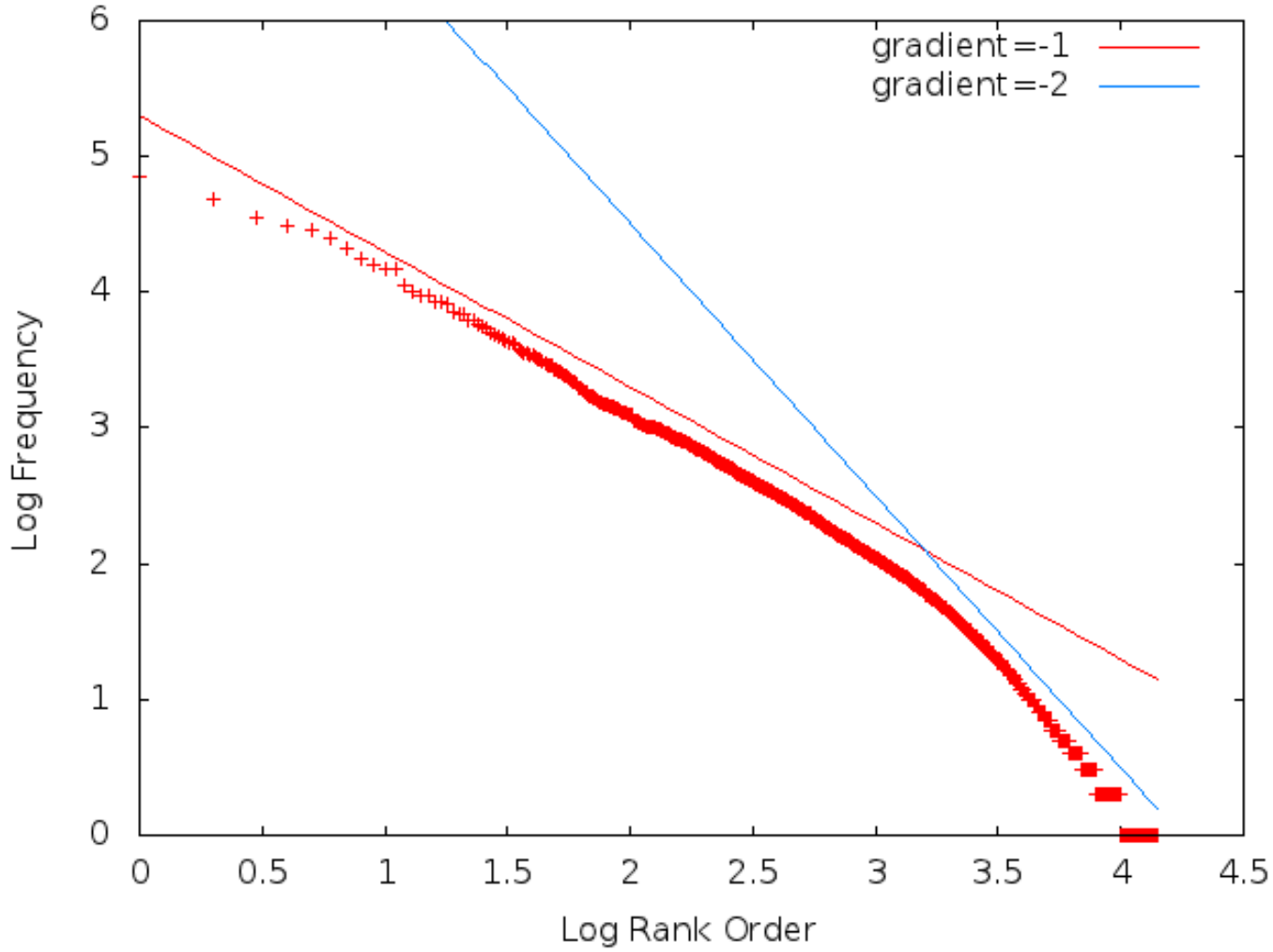
- We also get a lognormal distribution of document vocabulary size.

Type to Token Ratio

Australian National University

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

Zipf Curve


gradient=-1
gradient=-2
Log Frequency
Log Rank Order
0
1
2
3
4
5
6
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

Token Occurrence

Australian National University

- We extracted the 500 most frequent terms:
  - The least most frequent term in this list appears 249 times.
  - The list captures 78% of the terms in the corpus.
- The frequency of these terms were also calculated in both Brown and Reuters and then compared.
- e.g The token 'or' appeared 20.1 per 1000 tokens in the contract corpus, whereas, it only appeared 1.9 per 1000 tokens in Reuters and 3.6 in Brown.
- 'any' appears far more frequently in the contract corpus.
- 'was', 'his' and 'it' all appear less frequently (it was expected).

Most Distinctive Terms

- Log likelihood measures were calculated in comparison with the Brown and Reuters corpus.

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	confidentialit

Key Term Measures

- Using domain knowledge the frequency of a short list of key terms were compared. They include:
  - 'if' – used to mark a conditionality
  - 'may' – marks freedom
  - 'must' – marks obligation
  - 'means' – marks a definition
  - 'where' - marks conditionality

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

Collocations

Australian National University

- 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; 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 also contains: numde solver; flight attendant; mobile phone; nrl club; and rugby league.

Chunk Analysis

Australian National University

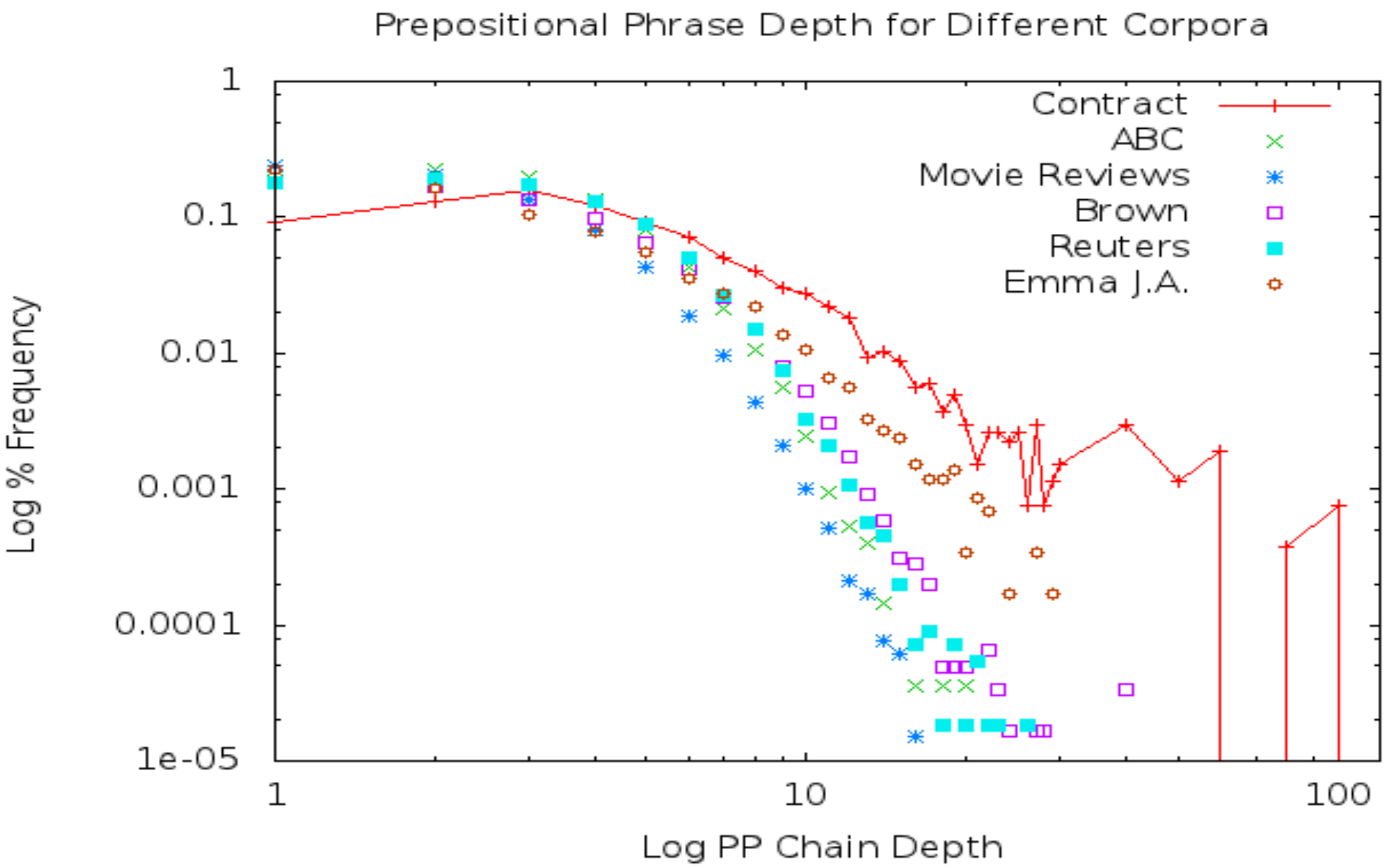
- Explored phrase occurrence in the contract corpus (comparison with related work by Venturi)
- Used a sub-corpus of 50 contracts. All non-rule material (such as headings, tables of contents, execution blocks) was stripped from this sub-corpus.
- MontyLingua was used to apply parts of speech tags and to chunk the sub-corpus.
- Comparison was made with:
  - Brown,
  - Reuters,
  - ABC rural reports,
  - ABC science reports,
  - Austen's Emma, and
  - Movie Reviews.

Chunk Analysis

Australian National University

- The contract corpus had comparatively long sentences. On close inspection this was due to the inclusion of long lists.
- The contract corpus had relatively low usage of verbs at a sentence level.
- In the contract corpus the occurrence of prepositional phrases was notably high. This was similar to the study by Venturi on environmental law.

Chunk Analysis



Conclusions

Australian National University

- We have reported the design and profiling and phrase analysis of a corpus of Australian contract language.
- The profiling of the corpus supports the validity of the method employed in compiling the corpus. (e.g. conformance with Zipf's law, a lognormal distribution of document length and vocabulary)
- We identified some distinctive terms used in contract language.
- We have provided chunk analysis of the contract corpus.
- The corpus provides a useful tool for future work on contracts.

Thank you.

Questions?

Chunk Occurrence

Australian National University

	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