Part 1 — Collect a bunch of contracts and study the language

Before you can use language as input for a software tool you need to know how that type of language works and what makes it different from language in general. So we collected around 250 contracts and studied them. This kind of work is called corpus linguistics.

We found:
- Our contracts re-used the same words more than three times as often as English in general. On average a contract has a vocabulary of about 630 words.
- Most contracts are going to be between 700 and 13 000 words long.
- Some simple function words such as ‘or’, ‘any’, ‘the’ show up more often in contracts than language in general.
- Words like organisation, GST, authorised, licence, provider, mediation, invoice, copyright, waiver, ABN, funding and confidentiality can help tell us we are looking at a contract.

That’s the good news. The bad news is contract sentences are long. Really long - and they have a sting in the tale. A sentence in a contract will have about 5.5 prepositions. Analysing such a sentence will not be as easy as analysing standard English.

See Michael Curtotti and Eric McCreath A Corpus of Australian Contract Language Proceedings of the 13th Annual Conference on Artificial Intelligence and the Law 2011 Pittsburgh

What if contracts helped write themselves?

Michael Curtotti ANU CECS IHCC Group, michael.curtotti@anu.edu.au.
Panel: Dr Eric McCreath (Supervisor), Associate Professor Chris Johnson (Chair), Professor Tom Gedeon, Dr James Popple

Part 2 — Sort the Sheep from the Goats

The important part of a contract are the ‘rules’ which describe what the parties have agreed. But the rules are surrounded by other matter we’re not interested in, like headings, appendices, tables of contents.

So how do we sort the sheep from the goats? Now that we know what makes contract language different we can see if some of the simple words which contracts use differently can help us do that. Applying machine learning we find that they do.

On a ‘head’ of 9000 paragraphs, such words can give us an accuracy of 88%. Combined with simple layout features we get an accuracy of 91%. That’s nine sheep for one goat.

Michael Curtotti and Eric McCreath Corpus Based Classification of Text in Australian Contracts Proceedings of 2010 Australasian Language Technology Association Workshop describes some of our classification work.

Part 3 – The Secret Life of Contracts

A graph (or a network) is a set of nodes with edges between them. Contracts have multiple graphs hidden in them. For example its common for contracts to have a list of definitions. When you read a rule you also need to read each definition used in the rule. Each definition in turn might refer to another definition. So how long is the rule really? What if the rule is too long? Can we really understand it?

Visualising how definitions are linked together help us understand what is happening in a contract.

On the left you see how one definition is used in another in a sample contract. At a glance you can tell that you have an agreement involving a ‘scholarship student’ for a ‘research studies project’. These two terms have the most incoming edges. You can see too that to know what ‘Confidential Information’ means, you also need to know what ‘New Material’, ‘Research Studies Project’, ‘Scholarship Student’, ‘Parties’ ‘Scholarship’ mean.

“New Material” means any Material that is created, written or otherwise brought into existence by the Scholarship Student and/or the Parties, in the process of carrying out the Research Studies Project.

See the goats? Now sort them out.