Smart Tools for Contract Drafting
Michael Curtotti*

1 Introduction
This poster reports initial research to develop novel, practical and unobtrusive software tools for drafting contracts. The research is being carried out on a part-time basis for an MPHil.

The research seeks to apply and develop knowledge and methods from the fields of machine learning and natural language processing for this purpose.

A secondary aim of the research is to carry out a comparison of the professional paradigms of software engineers and contract lawyers.

Industrial need
Contracts are found everywhere. They govern economic transactions across society. Contract drafting stretches back to the beginning of recorded history.

Contracts are found everywhere. They govern economic transactions across society. Contract drafting stretches back to the beginning of recorded history.

The primary tool used in modern contract drafting is a word processor. Tools to assist drafters would assist them to more accurately and quickly draft contracts. Ambiguity in contracts is hard to detect and can impose substantial costs on parties to a contract (see figure 3). Software tools can complement the skills of the human drafter in identifying and resolving ambiguity. An outcome of this research will be the creation of tools enhancing word processors which are designed for the needs of contract drafters.

Prerequisites for the creation of drafting tools
The creation of tools for contract drafting requires knowledge of how lawyers draft, of contracts as a variety of written English and of natural language processing as an tool for analysis of the textual content of contract documents. Figure 2 shows a prototype of a tool which identifies defined terms in a contract and allows a drafter to access its meaning by rolling over the term.

2 Coming to grips with Legal English
A common lay view is that legal English is impenetrable. The difficulties of legal writing have given birth to the ‘plain English’ movement which aims to make legal texts more accessible to their users.

Empirical analysis of language use in legal documents, particularly in contracts, is however rare. This research has resulted in the creation of a ‘corpus’ (or electronic collection) of 256 contracts; or around 1,000,000 words of text.

By comparing this corpus against representative collections of general English (e.g., the 1,000,000 word Brown corpus), we can begin to characterise how contract English is different to general written.

For example, words such as ‘must’ and ‘may’ play a special role in contracts: often used to express obligations and powers, respectively. The words ‘means’ and ‘includes’ are the primary markers of defined terms.

However such terms are used in contracts as compared with English generally? Figures 4 and 5 explore some differences in language use.

3 Classification of Contract Data
Natural language processing typically involves a series of operations on a text with the ultimate aim of information extraction. Initial work as part of this research has focussed on segmenting contracts into ‘lines’ or ‘paragraphs’ and undertaking multi-class labelling of that data.

Examples of labels applied include “clause matter”, “definitions”, “clause headings”, “schedule material” etc.

High accuracy has been demonstrated in labelling with average labelling accuracy of 94.1% using a random forest machine learning algorithm (with accuracy as high as 97% achieved for key classes).

In a jointly authored paper submitted to a conference: we have also shown that a useful strategy is to combine machine learning and rule based methods to increase accuracy. Figure 6 shows that the combined method is particularly useful where only a few examples of a class are available (a scenario where machine learning is particularly weak).

The following shows “corrections” applied by a decision tree machine learning algorithm to hand coded tagger output. The capitalized terms are the class labels. The lower case terms are the logical tests applied by the decision tree in the labelling process. E.g. if the number of tokens are less than 4 and the line is further less than 0.38 through the contract, change the label to ‘PRELIM’.

4 Conclusion
Initial work has begun to outline the nature of Australian contract language and has demonstrated high accuracy in classification of line data from Australian contracts, a useful step for later natural language processing.

An initial prototype has been developed for a definition tool with the following functions:

- automatic identification of definitions;
- highlighting of definitions;
- hyperlinking occurs with links back to the definition;
- automatic pop up of the definition meaning;
- once click removal of definition highlighting.

The next phase of work will focus on a systematic statistical characterising Australian contract language.

* Information and Human Centred Computing Group, College of Engineering & Computer Science, Australian National University
* Jointly authored with Dr Dic McCreath