**Enhancing the Communication of Law:**
**a cross-disciplinary investigation applying information**
**technology**

**Michael Angelo Curtotti**

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

October 2016

Except where otherwise indicated, this thesis is my own original work. This thesis is undertaken by compilation and includes seven papers annexed in the Appendices. Except as stated below, I substantially wrote each of those publications and carried out the bulk of the research associated with each of them. In respect of the paper attached as Appendix A.3 (*Interdisciplinary Cooperation in Legal Design and Communication*), I wrote over 75% of the paper. Other than literature research, the bulk of which I carried out, that paper did not involve other research. The total word count (including published papers annexed) is less than 99180, excluding abstracts, acknowledgements, tables, appendices, footnotes and bibliographies.

Michael Angelo Curtotti
10 October 2016

To Ranjana,
To whose love and support, this work is due.

# Acknowledgments

A PhD, as many have discovered, is a demanding and life changing journey. It is a journey that no one undertakes or can complete alone. Many assist in the process and without their contribution, encouragement, wise counsel and support, the journey would never be complete.

My family deserve the largest measure of thanks, particularly my wife Ranjana, for their unstinting support and love over the years that this thesis has taken.

Dr Eric McCreath, my supervisor, has been an unfailing companion through this journey and I thank him. I have often had cause to feel grateful for his support and guidance. Not only has he been a wise mentor in this journey, he has been a collaborator in most of the research papers that form part of this thesis. He has been unfailingly patient and supportive in his guidance and the journey would have faltered at many points but for his support. In those moments of internal doubt, such as many PhD students experience, when the journey might have been abandoned, his kind and scholarly spirit has been the final irrefutable argument for the case that the journey must be completed. But for his mentorship, the journey would not have reached its end. Dr McCreath was also one of my first teachers as an undergraduate student undertaking computer science units (for a never completed Bachelor of Science). It was in those lectures that a love and fascination for computer science was born. Those first revelations, deepened in other computer science units, became the food for thought that have ultimately led to this point.

I would also like to thank the other members of my supervisory panel, Associate Professor Chris Johnson (Chair), Professor Tom Gedeon and Dr James Popple for their unstinting generosity, support and guidance. The regular meetings throughout the research and the thoughts, suggestions, re-focussing as well as revisions they have kindly proposed, have been invaluable in guiding the research in productive directions and improving its outcomes.

One of the pleasures of having undertaken this journey as thesis by publication has been the opportunity to work with fine co-authors who have contributed their own measure to the work. Dr Helena Haapio and Ms Stefania Passera, for their collaboration in the field of visualization; Mr Tom Bruce, Ms Sara Frug, Mr Wayne Weibel and Mr Nic Ceynowa of the Cornell University Law School Legal Information Institute for their collaboration in readability research; Dr Srinivas Sridharan for col-

laboration in network analysis and visualization and Dr Eric McCreath throughout. I thank them for their collegial support and the pleasure that comes from working together with enthusiastic collaborators who share with you a love of knowledge and a desire to explore its boundaries. I would also like to thank my co-authors for their kind approval for inclusion of the attached papers in this thesis.

Also to be thanked is my my friend since school days, Dr Peter Spooner, for his support with insightful materials on the PhD process and kindness in reviewing two of the papers included in this thesis.

The ANU Research School of Computer Science also deserves thanks. Without its support for travel it would not have been possible to present this research at a number of international conferences. These conferences and the invaluable feedback from anonymous reviewers all contributed to enhancing the quality of the research. As a part-time research student I have been largely an invisible presence. I have nonetheless felt a part of a research community for which I have a great deal of affection. It is a community characterised by a spirit of innovation, an openness to new ideas and an absence of 'orthodoxies' which implicitly limit thinking or investigation. It is this, in my view, which is the little known secret of the success of computer scientists in changing our world. They have given us new ways of thinking. The machines with which we are now familiar, "computers", are but the physical shadows of the reflections of their minds.

I would also like to acknowledge the free and open source movement. At virtually every point of this research, its progress has depended on the generosity of those who consider it more important to make knowledge available to all, rather than to seek direct financial advantage from it. The availability of data and software to process that data such as the Natural Language Toolkit, Weka Machine Learning Software, Python, the R-statistical package and others have all contributed to this research.

Finally, I would like to thank those anonymous citizens who participated in the citizen science project reported in this research.

# Abstract

Law is pervasive in culture. It is a form of communication between government and citizens. When effective, it is a tool of government policy. If poorly designed, law results in unnecessary costs to society. Impediments to understanding of the law limits and distorts democratic participation. Yet, historically, the law has been inaccessible to most. Thus enhancing the communication of law is an important and standing problem. Much work has been done (for example through the plain language movement) to improve the communication of law. Nonetheless, the law remains largely unreadable to non-legal users. This thesis applies information technology to investigate and enhance the communication of law. To this end, this thesis focusses on four main areas.

To improve the readability of law, it must be better described as a form of language. Corpus linguistics is applied for this purpose. A linguistic description of contract language arose from this work, which, along with the corpus itself, has been made available to the research community. The thesis also describes work for the automatic classification of text in legal contracts by legal function.

Reliable measures for the readability of law are needed, but they do not exist. To develop such measures, gold standard data is needed to evaluate possible measures. To create this gold standard data, the research engaged citizen scientists, in the form of the online "crowd". However, methods for creating and using such user assessments for readability are rudimentary. The research therefore investigated, developed and applied a number of methods for collecting user ratings of readability in an online environment. Also, the research applied machine learning to investigate and identify linguistic factors that are specifically associated with language difficulty of legislative sentences. This resulted in recommendations for improving legislative readability.

A parallel line of investigation concerned the application of visualization to enhance the communication of law. Visualization engages human visual perception and its parallel processing capacities for the communication of law. The research applied computational tools: natural language processing, graph characteristics and data driven algorithms. It resulted in prototype tools for automatically visualizing definition networks and automating the visualization of selected contract clauses.

Also, the work has fostered an investigation of the nature of law itself. A "law as" framework is used to query the nature of law and illuminate law in new ways.

The framework is re-assessed as a tool for the experimental investigation of law. This results in an enhanced description of law, applying a number of investigatory frames: law; communication; document; information; computation; design and complex systems theory. It also provides a contrastive study with traditional theories of law - demonstrating how traditional theories can be extended in the light of these multidisciplinary results.

In sum, this thesis reports a body of work advancing the existing knowledge base and state of the art in respect of application of computational techniques to enhancing the communication of law.

# Contents

# List of Publications

## A. Publications Included in this Thesis

**A.1.** Curtotti, M., McCreath, E., Bruce, T., Frug, S., Weibel, W. and Ceynowa, N. *Machine Learning for Readability of Legislative Sentences*. Proceedings of the Fifteenth International Conference on Artificial Intelligence and the Law. (ICAIL, 2015) Association for Computational Machinery (ACM) [In Press]

**A.2.** Curtotti, M., Weibel, W., McCreath, E., Ceynowa, N., Frug, S. and Bruce, T. *Citizen Science for Citizen Access to Law*. Journal of Open Access to Law, 2015, Vol 3:1. [In Press]

**A.3.** Curtotti, M., Haapio, H. and Passera, S. *Interdisciplinary Cooperation in Legal Design and Communication*. Kooperation - Digitale Aasgabe zum Tagangsband des 18. Internationalen Rechtsinformatik Symposions IRIS 2015 (Cooperation - Digital Proceedings of 18th International Legal Informatics Symposium IRIS 2015), JurisletterIT 26 February 2015. [In Press]

**A.4** Curtotti M. and McCreath E. *A Right to Access Implies A Right to Know: An Open Online Platform for Research on the Readability of Law*. Journal of Open Access to Law, 2013. Vol 1:1 [In Press]

**A.5** Curtotti, M., McCreath, E. and Sridharan, S. *Software Tools for the Visualization of Definition Networks in Legal Contracts*. (Research Abstract) Proceedings of the Fourteenth International Conference on Artificial Intelligence and the Law (ICAIL, 2013). Association for Computational Machinery (ACM). [In Press]

**A.6** Curtotti, M. and McCreath, E. *Enhancing the Visualization of Law*. Presented October 2012 20th Anniversary Conference of Law via the Internet - Cornell University [Peer reviewed, In Press] - publication via conference webpage presentation page at https://blog.law.cornell.edu/lvi2012/presentation/enhancing-the-visualization-of-law/ with link to SSRN page http://ssrn.com/abstract=2160614

**A.7** Curtotti, M. and McCreath, E. *A Corpus of Australian Contract Language*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law (ICAIL, 2011). Association for Computational Machinery (ACM). [In Press]

## B. Related Publications

**B.1** Passera, S., Haapio, H. and Curtotti, M. *Making the Meaning of Contracts Visible - Automating Contract Visualization*. In *Transparency. Proceedings of the 17th*

# List of Figures

# Introduction

*"Yes, very different. But I think Mr. Darcy improves upon acquaintance."*

*"Indeed!" cried Mr. Wickham with a look which did not escape her."*

Jane Austin, *Pride and Prejudice.*

Our relationship with the law is asymmetrical. It is a pervasive influence in our lives, communities and societies. Yet for countless individuals, and sometimes for entire demographics, effective access to the law is limited. How law is communicated profoundly affects its accessibility. It is for this reason that for centuries the language of the law has been contested and critiqued.

Recent decades have seen a wealth of work addressing the accessibility of law. One stream of work sought to ensure that the law would be readily available on the internet. The Free Access to Law Movement discussed in following pages is one aspect of this stream. A second stream concerns access to law in terms of its content: that its meaning should be accessible to its users. This stream of work is captured in the work of the plain language movement, which has had an extensive influence on how the law is communicated. Despite such developments, access to law remains profoundly impeded. Before the world wide web, the law was published in statute books and case books. Apart from professionals, whose task it was to engage with such material, few individuals would have attempted to find it, let alone read it. As the research reported below describes, vast new audiences now actually engage with the law, irrespective of level of professional training. However, as noted by one authority, having found the law, a member of that audience is likely to be little wiser.

*"Legislation affects us all. And increasingly, legislation is being searched for, read and used by a broad range of people. It is no longer confined to professional libraries; websites like legislation.gov.uk have made it accessible to everyone. So*

1

*the digital age has made it easier for people to find the law of the land; but once they have found it, they may be baffled. The law is regarded by its users as intricate and intimidating.*"[1]

The research reported in this thesis is concerned with applying information technology to enhance the communication of law. It is particularly focussed on the communication of legal "rules" in legislation and in contracts. It investigates applying information technology to this end and develops methods and tools (including computational tools) towards this goal.
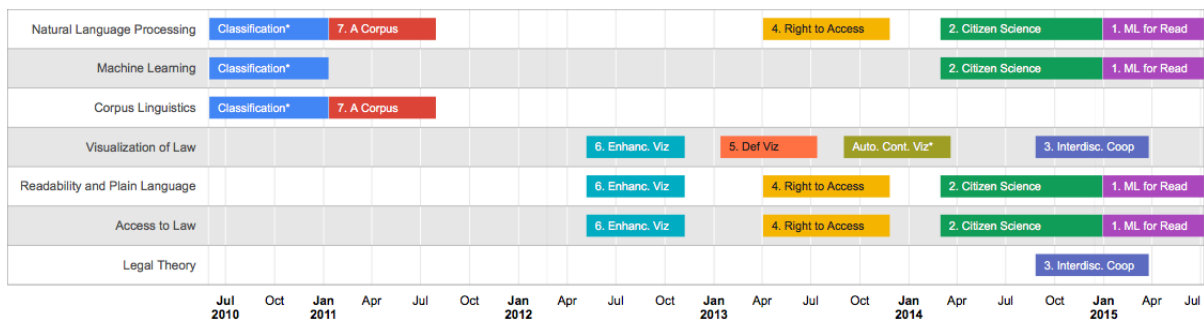


Figure 1.1: Thematic Timeline of Publications

The research is primarily contained in the seven published papers attached in the appendices to this thesis. Figure 1.1 provides a chronological overview of the papers included in the List of Publications found on page *xiv*. Each paper is thematically coded to summarise the primary areas the paper addresses. The two papers marked with an asterisk are not included with this thesis, for reasons of length, but are shown to illustrate their place in the overall flow of the research. The research process described in this thesis is, however, better understood as a single ongoing investigation into enhancing the communication of law. Given this the following discourse does not draw a sharp line between those papers which are annexed and those which are not. The reader may, where needed, refer to the appendices, where the included papers are annexed.[2]

In addition to the papers themselves, the main body of this thesis undertakes an additional substantial investigation into the nature of law which is further described in Section 1.1, below.

---

[1]The UK Parliamentary Counsel's Office Good Law Initiative, 2013. Cited in [Curtotti et al., 2015c].

[2]The two additional papers not included in this thesis may be accessed at the online locations indicated in the List of Publications. Note that Stefania Passera is the primary author of *Making the Meaning of Contracts Visible - Automating Contract Visualization*. In respect of that paper, I was the sole designer and implementer code for the automation of the contract visualization described in that paper. The visualizations are based on static designs developed by Stefania Passera.

The application of computational techniques reported in this research is undertaken in a multidisciplinary context. It engages legal, linguistic, computational, informatics and design paradigms. Further, the goal of the research fits within the broader goal of enhancing 'free access to law'; which provides a broader purposive frame. Inevitably, this brief description glosses over many complexities that arise. What, for example, do we mean by law? What forms of communication are encompassed within the goal? What measures may tell us that communication is being enhanced? What do we mean by application of computational techniques? And what is meant by open access? The research reported in this thesis explores these and related questions.

Generally, this work is addressed to written legal 'rules' - if we may for the moment limit our model of law to a "set of rules".[3] By legal rule, what is meant is a legal provision, a common form of which is a more complex version of:

> *If A, B (must / must not) do C.*

That is, by law is meant a concrete manifestation of written regulatory language found in a legal instrument such as a law of Parliament, an executive regulation or a commercial contract. This is, of course, a subset of phenomena to which the term "law" is applied. It does not, for example, include the written materials produced by judicial bodies, such as reported judgements, which are also part of "the law". Nor does it include written communication between lawyer and client, or documents produced by law enforcement bodies, which are also within the bounds of "the law".

The goal of promoting free access to law is a central concern of the Free Access to Law Movement (FALM). The movement has been highly influential in shaping how law is made available in online environments. FALM emerged as a collaboration of lawyers, software professionals and publishers. FALM was successful both in setting ground rules for legal materials in online environments, and in undertaking the practical work required to make legal materials available online. It emerged organically, in response to the potential of the world wide web for enabling an enhanced communication of law.[Martin, 1999; Martin; Greenleaf, 2010; Greenleaf et al., 2013] Free access to law thus forms an appropriate context in which to examine the further application of computational technologies to the same ends. The goal of this research falls within the broader goal of enhancing access to law.

While regulatory language is typically largely communicated in the form of words, visual communication of information found in regulatory language and its enhancement is also explored. Communication as addressed in this thesis, thus extends to

---

[3]The framing of law as a set of rules is here used as a linguistic and conceptual convenience. Section 2.1 of Chapter 2 undertakes an enquiry into the nature of law.

communication in both word and image. This aspect of the thesis is particularly concerned with visualization of legal rules, an alternative form of communication of law. Visual communication draws on concepts such as those explored in the legal design movement. In common with FALM, legal design is concerned with the communication of law, but in contrast to FALM, it emphasises the design discipline as its frame of reference. While the legal design movement may use computational technologies, it is not confined within the boundaries of a particular technology.

The thesis also applies computational linguistics (more specifically corpus linguistics) to study law as language. Machine learning is applied in the research to classify regulatory language according to its legal functions. Machine learning is also applied to investigate the readability of regulatory language. Crowdsourced methods in an online context, are used to collect data contributing to assessing the communication of law (i.e. its readability). Network approaches are employed to represent and analyse definitional networks in legal contracts, and provided a basis for visualization. Programs are created to automate the visualization of selected contract clauses. The online visualization of law is studied to assess the state of the art on its online presentation.

## 1.1 The Contributions Made by this Research

The research reported in this thesis makes a multifacetted contribution to enhancing the communication of law - particularly through the application of computational tools.

There is an existing and extensive body of work which approaches law as language, for example Mellinkoff's seminal work *The Language of the Law*.[Mellinkoff, 1963] Little of that work is however specifically concerned with the study of legal rules in contracts as corpora of linguistic data.[4] This thesis contributes to an understanding of the linguistic characteristics of regulatory language in contracts. It involved the production of the second publicly available corpus of contracts, as far as known to the author. The research analyses the corpus and discusses its characteristics. This work contributes to an understanding of law as a form of language. Investigations are carried out on the classification of text in legal contracts by their legal function. Machine learning and hybrid methods are applied to attain high accuracy in classification tasks on a set of test contracts. This was the first work of its kind applying hybrid methods to the automatic multi-label classification of the functional role of text in contracts. Only one prior work was identified which ad-

---

[4]See further Section 2 of *A Corpus of Australian Contract Language* for related work [Curtotti and McCreath, 2011] (Appendix A.7).

dressed functional classification of text in contracts - which limited itself to clause vs. non-clause classification.[Indukuri and Krishna, 2010],[Curtotti and McCreath, 2010] The work established the feasibility of such multi-label classification and proposed methods for optimising accuracy using hybrid methods. These methods are potentially valuable as inputs to further processing of contract language, and insights and rule based methods from this research are applied in visualization work described below.

Readability of law in general is a heavily studied problem - captured in the phrase "the plain language movement". Plain language began to influence the writing of legal rules in contracts by the 1960s.[Friman, 1994] By the 1990s executive orders mandated plain language for all regulations in the United States.[DuBay, 2004] Plain language imperatives provide widely accepted guidelines for the writing of legislation.[Kimble, 1994] Despite extensive evidence that plain language improves comprehension,[Benson, 1984] the goals of the plain language movement have not been achieved in respect of regulatory language. Studies of legislation, including legislation which has been revised in accordance with plain language guidelines leads to a conclusion that it remains very difficult to incomprehensible for most audiences.[GLPi and Smolenka, 2000; Smith and Richardson, 1999; Sawyer, 2010; Tanner, 2002][5] Readability metrics are a primary tool for measuring the readability of text.[DuBay, 2004] However this tool is neither designed for, nor is it suited for measuring the readability of regulatory language.[Woods et al., 1998; Melham, 1993][6] An increasing body of work has applied natural language processing and machine learning to the readability of text.[Collins-Thompson, 2014] However, no previous work has applied these techniques for readability of legislation.[7] This research contributes to the study of the readability of legislation in a number of ways. The state of the art in respect of the readability of legislation is collated and summarised and the current challenges of legislative readability identified based on the research literature. Research tools were created for extracting linguistic characteristics and information from text and were made available online. These tools were applied in a collaboration between the Cornell Law School Legal Information Institute and the ANU Research School of Computer Science to collect data on the users of legislation online, and to enable a citizen science project to assess the readability of legislative sentences. Data was collected over a three month period resulting in tens of thousands of data points.

---

[5]See further discussion Section 2.3 of *Citizen Science for Access to Law* [Curtotti et al., 2015c] (Appendix A.2).

[6]See further discussion in Section 2.1 of *A Right to Access Implies A Right to Know: An Open Online Platform for Research on the Readability of Law* [Curtotti and McCreath, 2013] (Appendix A.4), and Section 2.2 of *Citizen Science for Access to Law* [Curtotti et al., 2015c] (Appendix A.2).

[7]See Section 2.5 of *Citizen Science for Access to Law* [Curtotti et al., 2015c] (Appendix A.2) and Section 2.1 of *Machine Learning for the Readability of Legislative Sentences* [Curtotti et al., 2015b] (Appendix A.1).

The results were then used to carry out machine learning and correlation studies for the identification of linguistic features associated with language difficulty in legislative documents. As far as is known, this was the largest such project ever carried out to measure the readability of legislative language, and to understand the audience that reads it. The work advances an understanding of the problem of legislative readability and progresses the development of specific computational tools and methods for enhancing the readability of legislation. It proposes a number of measures for detecting difficulty of legislative language, based on results of the research.

A second primary direction of the research was concerned with the investigation of visual communication to enhance the communication of law. This work included application of natural language processing and networks to automatically extract and visualise definition networks in legal contracts. In [Curtotti and McCreath, 2012][8] primary research is carried out mapping the state of visualization of legislation in selected English speaking jurisdictions, particularly in online environments. As far as the authors are aware, this was the first such investigation of its kind. This review (which included both official and unofficial sites) found a diversity of approaches to visualization of law online. With the exception of novel and experimental visualizations most visualizations presented law as close digital analogues of documentary originals. There is of course extensive work in visualization more generally. The existing research literature discusses both the visualization of information (drawn primarily from computational fields) as well as the visualization of knowledge (which is not necessarily bounded within computational approaches).[Burkhard, 2004; Card and Mackinlay, 1997; Lengler and Eppler, 2007; Swienty and Takatsuka, 2010] None of this work is specifically concerned with the visualization of regulatory language per se. Grinstein proposes a conjecture for a visualization framework.[Grinstein, 2010] In the context of an absence of systematic evaluation of online visualizations of law, the research reported in this thesis adapts this conjecture to propose a revised model as a framework for evaluating and enhancing online visualization of legislation.[9]

In respect of visualization of contracts, the research reported here advances the automatic visualization of information from legal contracts. Natural language processing and network methods and graphical tools are used to create four different prototype visualizations of definition networks in contracts, with different use cases. Also in the field of visualization, this thesis reports multidisciplinary collaboration carried out with researchers in Europe to automate the visualization of the *meaning*

---

[8]See Section 2 of *Enhancing the Visualization of Law* (Appendix A.6).

[9]See further Section 4 of *Enhancing the Visualization of Law* [Curtotti and McCreath, 2012] (Appendix A.6).

of selected rules in business to business contracts. The research literature in this field includes natural language processing to extract definitions including in legislation (although not in respect of contracts).[de Maat and Winkels, 2008; Degórski et al., 2008; Winkels and Hoekstra, 2013] It also includes graph visualization techniques,[Feldman and Sanger, 2007] generation of word clouds,[Halvey and Keane, 2007; Lohmann et al., 2009] and the development of logical languages for representation of contract rules.[Daskalopulu and Sergot, 1997; Governatori, 2005] The research reported here resulted in novel visualizations of definition networks in legal contracts and proof of concept of automated visualization of selected contract clauses.

In addition to published results the research lead to a number of practical results, including the creation of legal corpora (an Australian Contract Corpus and a Corpus of American Regulatory English labelled for readability). Other practical outcomes included a Readability Research Platform made available to researchers, a prototype tool for visualization of contract clauses and an online tool for visualization of definition networks in legal contracts. These practical outcomes are discussed in Chaper 3 of this thesis.

Finally, the research as a whole led to a process of reflection and investigation as to the nature of law. This investigation is embodied in Section 2.1 of this thesis. From a multidisciplinary perspective, it asks the question "What is Law?" With few exceptions, the primary thrust of existing theories of law conceptualise or problematise law conceiving it primarily as a set of rules.[William, 1980],[Freeman, 2001], [Hart, 1961][Penner et al., 2002],[Wacks, 1999],[Doherty, 2002],[Letwin, 2005] (See further discussion in Section 2.1.2 of this thesis). The research reported here applies a "law as ..." paradigm for exploration of this question.[Lavi, 2010],[Mertz and Rajah, 2014],[Tomlins and Comaroff, 2011],[Haapio, 2013, pp27-40] This paradigm is investigated and re-conceptualised as a method for undertaking 'thought experiments' on the nature of law. As a result, this research departs from truth claims both of traditional theories and some of the existing work applying a "law as" paradigm. The investigation hypothesises and presents a fuller account of the nature of law, synthesising the results of these thought experiments into a multifacetted understanding of the nature of law. In doing so, it presents an alternative to reductive, analytical approaches more traditionally applied in seeking to understand the nature of law.

The research has a number of limitations which are further discussed in Chapter 4. Among these are a limitation to the English language context and limitation as to legal jurisdiction (primarily addressing legal instruments from Australian and United States jurisdictions). Areas where the research could be further extended in future include the expansion of gold standard labelled readability data for legislation, the extension of citizen science methods in application to the readability of legisla-

tion and the development of methods to investigate the communication of legislative *knowledge*, as opposed to legislative information. (See further discussion in Section 4.2.2.)

## 1.2   Outline

This chapter introduces the thesis and its contribution.

Chapter 2 provides a theoretical background to this thesis. It examines the major theoretical contexts for the reported research. These theoretical contexts are legal informatics, access to law, readability, visual communication and design. However before addressing these topics the chapter investigates the nature of law. Asking, from an interdisciplinary viewpoint, the question: what is law? In answering this question, the insights of the reported research are drawn on.

Chapter 3 focusses on the experimental methods of the research. It describes the data which is subject of investigation and the outputs in which it is embodied (for example legislative corpora). The experimental tools used in the research are described. These tools included natural language processing, corpus linguistics and machine learning. The software tools used or created are also enumerated and the methods of dissemination reported.

Chapter 4 provides a concluding review and synthesis of the work reported in this thesis. The first section of the chapter reviews each attached paper in turn; highlighting its contribution in the context of the existing research literature. The second half of the chapter provides overall conclusions of the research, including a synthesis of the contribution of the research. It discusses limitations of the research and identifies potential future research directions.

Chapter 5 provides abstracts as well as visual summaries for each paper in the form of 'word clouds'.

Following the Bibliography, the appendices provide copies of the published papers compiled with this thesis.

# A Theoretical Context and Investigation

The purpose of this chapter is to provide the theoretical context of this thesis. Six theoretical frames are addressed. Firstly, any investigation of law, raises implicitly, if not explicitly, questions as to the nature of law. This is particularly true in the case of a multidisciplinary investigation of law, which approaches law through lenses beyond the legal paradigm. Accordingly, Section 2.1 discusses the nature of law. Second, the application of computational techniques to law particularly engages the field of legal informatics, which is centrally concerned with application of computers to the legal field in general (Section 2.2.1). Third, enhancing the communication of law, falls within the broader theoretical and practical frame of enhancing access to law (Section 2.2.2). Fourth, the issue of readability in general and readability of law in particular, has its own literature and background which is one of the primary limbs of the work reported in this thesis (Section 2.2.3). Fifth, visualization of law falls within the frame of visual communication, which again has its own theoretical frame (Section 2.2.5). Sixth, and finally, a design paradigm is engaged through multidisciplinary collaboration in visual communication, engaging a further theoretical frame. Each frame draws out different insights and taken together they provide a multifacetted theoretical framework.

In one respect, this chapter provides a review of relevant background to the publications compiled with this thesis. Accordingly, extensive reference is made to related theoretical literature which supports this contextual background. In another sense, this chapter uses the published papers as a departure point for further investigations. In this respect, this chapter serves in part as exegesis on the published papers. The papers are not described in this chapter, rather pointers in their content are retrieved and further explored. Sometimes these pointers have been consciously placed in a paper with the thought that they offered potential for further exploration: for exam-

ple, a discussion of the nature of law is introduced in [Curtotti et al., 2015a],[1] and is here substantially extended in Section 2.1. [Curtotti et al., 2015c][2] outlines what might be meant by "access to law" and this theme is further explored in Section 2.2.2 and complemented with a discussion of the historical evolution of the concept in the Free Access to Law Movement. In other cases, particular work is examined for its theoretical implications, which were not discussed, nor necessarily envisaged, in the publication itself. For example "law as network" discussed below in Section 2.1.6 references [Curtotti et al., 2013],[3] which reports work on the development of software tools to visualize definition networks in contracts. That paper did not consider the broader theoretical implications of approaching legal statements as networks. Nonetheless such implications invite exploration by their presence in the published papers. This chapter thus looks both backwards at the existing literature and theoretical frames, including the published papers; and forwards in drawing out further theoretical implications.

The discussion of the nature of law which follows begins by reviewing the core of existing theories on the nature of law. This is followed by the application of a "law as ..." framework to investigate the nature of law from a multidisciplinary perspective. The "Law as ..." approach is outlined in Subsection 2.1.2.6. It uses "metaphor" as a tool for open-ended investigation of law. Thus, in this context, "law as document", "law as computation" and "law and design" are instances of such investigation. Essentially, what do we learn about law if we think of its documentary characterstics, or as a process of design? The approach is understood as a series of thought-experiments - each contributing cumulative insights on the nature of law. The answer to the question "What is Law?" arises as a synthesis of the insights offered by each investigation and existing theory. As a result, it is possible to conclude that law is far more than a "set of rules". Rather the "rule" characteristics of law forms a small subset of an ontology of law - that more fully describes its character. Further through consideration of law as a "complex adaptive system", it is concluded that reductive definition of law (again as is traditional) substantially impoverishes the proper description of the phenomenon of law. The very complexity of law requires such fuller description if an adequate level of accuracy as to its nature is to be maintained.

---

[1]See Section 6 of *Interdisciplinary Cooperation in Legal Design and Communication* (Appendix A.3).
[2]See Section 2.1 of *Citizen Science for Citizen Access to Law* (Appendix A.2).
[3]*Software Tools for the Visualization of Definition Networks in Legal Contracts* (Appendix A.5)

## 2.1   On the Nature of Law - a Multidisciplinary Investigation

*"Think not that We have revealed unto you a mere code of laws. Nay, rather, We have unsealed the choice Wine ..."*

Bahá'u'lláh, Kitab-i-Aqdas, paras 4, 5.[4]

### 2.1.1   Motivation

What is law? This question is central to the field of legal theory and it has attracted extensive investigation by legal scholars. The question has given rise to a diversity of perspectives, which are often interpreted as competing and mutually exclusive ways of understanding the law.

Why address this question in what is primarily an application of computational technology to the law? Firstly, it illuminates the research itself. How has the law been explicitly (and implicitly) understood in this investigation? Secondly, how does our understanding of law change as a result of such a multidisciplinary journey? To fail to investigate that change is to neglect important insights that multidisciplinary research inherently provides. These reflections on the nature of law are not descriptive, but rather extrapolate from the results and experience of the body of research with which this thesis is concerned. They motivate understandings of law that arise from and respond to engaging with law in novel and multidisciplinary ways.

Nonetheless, investigation of the nature of law is a crowded theoretical field. Might a perspective primarily informed by computer science and informatics have anything to offer this already well-developed discourse? Twining's insights on the contribution of multidisciplinary investigation to jurisprudence provides an intuition that it may do so:

> *"the main function of the jurist is as a conduit. [The jurist] ventures forth from the law to garner what one or more neighbouring disciplines have to offer respecting question of a general nature that have been thrown up in legal contexts. [The] role is to bring back the ideas, techniques, and insights of that other discipline and to integrate or assimilate them into the intellectual milieu of the law."*[Twining, 2001, p 21]

Section 2.1 undertakes such a multidisciplinary integration. As already mentioned, this theoretical enquiry arises from research applying informatics to the law. This research both required and invited an engagement with law that stepped beyond most theories as to the nature of law. They required an exploration of law in

---

[4]http://reference.bahai.org/en/t/b/KA/ka-4.html#pg20 accessed 13 November 2015

non-traditional ways because of the outcomes sought by the research. For example, examining the readability of law as done in [Curtotti et al., 2015b],[Curtotti et al., 2015c][5] and [Curtotti and McCreath, 2013][6] or investigating the linguistic characteristics of a body of legal texts, as done in [Curtotti and McCreath, 2011][7] requires the law to be conceived of and approached as a body of language and a process of communication. Undertaking machine learning on legal texts, as reported in [Curtotti et al., 2015b][8] and [Curtotti and McCreath, 2010][9] interprets the law as a body of data within which patterns are embedded. Visualization of law as reviewed and demonstrated in [Curtotti et al., 2015a],[10] [Curtotti et al., 2013][11] and [Curtotti and McCreath, 2012][12] views law as a form of textual communication which can be re-communicated using visual cognition. A focus on usability and user experience are invited by a focus on the design, as reported in [Curtotti et al., 2015a][13] and [Passera et al., 2014][14]. Such investigations invite a further examination of the nature of law, because they open perspectives on the nature of law that have been little explored in the past. Law as data, law as network, law as designed artefact are all suggested by the computationally oriented investigations listed above. From a practical viewpoint of the projects concerned, failing to investigate received assumptions as to the nature of law closes off potentially fruitful avenues of exploration.

What will become evident as perspectives on the nature of law suggested by a computational orientation are explored, is that these perspectives open up new ways in which the law can be put to use. A further justification for this investigation is thus practical. For example, without a reconceptualisation of law as data, the provision of access to law is significantly hampered. Further, these new perspectives throw new light on existing theories, which enable them to be approached in new ways.

There seem to be few works that explicitly explore the specific connections between informatics and how law is conceptualized. Aguiló-Regla does so, from a viewpoint of a sometime participant in legal informatics.[Aguiló-Regla, 2005] He contrasts two time points in legal informatics: its early days, in the 1980s and 1990s, during which expert systems were dominant, and the current era. He observes that in the intervening period the concept of law in civil law countries has moved away

---

[5]Appendix A.2.

[6]Appendix A.4

[7]Appendix A.7

[8]Appendix A.1

[9]*Corpus Based Classification of Text in Australian Contracts* http://ssrn.com/abstract=1885490

[10]Appendix A.3

[11]Appendix A.5

[12]Appendix A.6

[13]Appendix A.3

[14]*Making the Meaning of Contracts Visible - Automating Contract Visualization* http://ssrn.com/abstract=2630609

from an 'objective' and deterministic view of law. In that view, law is a system of rules that may be applied to specified cases to obtain a result - whether the result is determined by the mandates of the rules, or in indeterminate cases, by judicial discretion (itself a product of a rule). In more recent years, 'rule' has been joined by 'principle' as part of a description of law. Values are also now relevant. The law is embedded and inter-related with politics and culture. People interact with the law in many ways. Law is a practice in which a variety of human actors engage. Rather than applying informatics to further query the nature of law however, Aguiló-Regla considers the implications of such newer views of law for legal informatics. These theoretical changes may to some degree begin to trace the limits of deterministic approaches to law - such as rule based expert systems. On the other hand, new vistas open up for legal informatics, which can be applied to meet the information needs of a diverse audience engaging with the law in diverse ways.[Aguiló-Regla, 2005]

Gordon discusses the links between the field of artificial intelligence and legal theory, suggesting that artificial intelligence and the law should be considered to be part of legal theory. In making this proposal, he suggests that informatics engagement with the law, should be informed by legal theory.[Gordon, 2008]

Such observations are interesting, but is it possible also to go in the opposite direction: to revise concepts of law using insights from legal informatics and associated fields?

Section 2.1.2 briefly introduces some theories of the nature of law. While only expressing this rich field in outline, it offers a background for considering the significance of the following sections which look at law from multidisciplinary perspectives.

### 2.1.2   Related work on the Nature of Law

*"...the end of law is not to abolish or restrain, but to preserve and enlarge freedom..."*

John Locke, Second Treatise of Government, Of Civil Government: Book
II, para 57.

This section summarises key schools of thought on the nature of law. With some variance, text after text deals with a number of core theories about the nature of law. Natural law, positivism and its variants, legal realism, economic or sociological approaches to law, critical legal theory, feminist legal theory, critical race theory and marxist legal theory being some of the common grist to the jurisprudential mill.[William, 1980],[Freeman, 2001], [Hart, 1961][Penner et al., 2002],[Wacks,

1999],[Doherty, 2002],[Letwin, 2005] Natural law, positivism, American legal realism, critical legal theory and feminist legal theory are outlined below.

### 2.1.2.1 Natural Law

As it is described in many texts on legal theory, natural law is the 'original' or oldest theory of law. In this theory, laws exist in the abstract. The law either arises from the will of a divine being or is an expression of nature. In both cases, law is immaterial, universal and eternal. Its content may be derived by reason, or by divine revelation. Natural law is thus a 'source' of higher law available to legal actors and citizens above and beyond the temporal regulations of a state or sovereign. Natural law theory leads to conclusions such as that a citizen need not obey a ruler whose laws are unjust, or who lacks proper authority. Of course such a theory has political implications and natural law has played a role in democratic revolutions and in the evolution of international law, including human rights law. Although tracing its origins back to stoic philosophy of the ancient Greek and Roman era, natural law still retains its proponents. For example J.M Finniss, who derives certain immutable principles of natural law based on reason, including absolute duties that correlate with absolute natural human rights.[William, 1980, Chapter 2],[Freeman, 2001, Chapter 3],[Hart, 1961, p 182]

### 2.1.2.2 The Command Theory of Law and Legal Positivism

The command theory of law together with its broader framework of legal positivism is often presented as a polar opposite to natural law. Where natural lawyers sought to 'discover' abstract pre-existent principles, positivists looked to the real world exercise of power. Laws or rules are thus a species of command in the Austinian command theory of law. In particular, laws are "general commands communicated by a recognised sovereign power which is habitually obeyed and which can punish disobedience".[Curtotti et al., 2015a][15] (See [Bix, 1999, p34], [William, 1980, pp24 et seq]) Virtues of this theory are said to be that it enables law to be distinguished from "non-law" (such as morality) and that it focusses on the centrality of legislation.[Doherty, 2002, pp-75-75] It has been elaborated and critiqued by positivist theorists such as Hart, whose views are discussed below.[Hart, 1961] Austin and Bentham (his precedessor in positivism) developed their theories at a time when scientific determinism was dominant and it was believed that the natural world itself was deterministic and fully predictable. Legal theories reflected this determinism.[Holz, 2007]

---

[15]*Interdisciplinary Cooperation in Legal Design and Communication* (Appendix A.3)

Kelsen's theory of pure law is a variation on a theme. It is pure, because it excludes "non-law" and understands law as norms (i.e. rules) which a legal person has a duty to obey; in the sense that failure to obey is a condition for official sanction. Kelsen considered his approach to be scientific. It is the existence of the norm of punishment that is central to Kelsen's approach and which makes a norm 'law'.[William, 1980, pp60,63] Thus, in Kelsen's scheme, a law can be reduced to a formula which states: norm plus breach implies sanction. Moreover, individual norms form part of a network of norms. In this network some norms have real world effect in specific situations and may result in sanctions if they are breached. They are, so to speak, the leaves of a mathematical graph forming a tree. The 'leaves' of this network have a hierarchical organisation back to a basic norm or norms which determine how norms are made (e.g. the Parliament makes laws). Outside this framework of networked norms, we depart the realm of law and enter realms such as politics. In other words, to ask why Parliament makes a law is to leave the realm of law.[Leiboff and Thomas, 2004, p101-104]

Hart, who is regarded by some as the 'leading contemporary legal philosopher', [Wacks, 1999, p59] continues the focus on rules. He develops a schema in which law is one kind of rule. Other kinds of rules include social rules or rules of morality. Hart does not rely on a sovereign to give law its character as such. Rather, he looks to a certain 'minimum content' of law (although positivist himself adopting here a natural law flavour). This minimum content arises from the physical and social reality of the human person. But law is not valid because of these realities, rather its existence is explained by these realities. Law is necessary because of the human condition. Generally accepted social practice is substituted for the sovereign as the source of law. Hart develops a more detailed theory of legal rules than found in Austin. He distinguishes primary rules (which impose obligations) from secondary rules. Secondary rules are in turn sub-divided into rules of change, adjudication and recognition. Rules of change, as their name suggests, provide powers to individuals and agencies for adapting (changing) primary rules of a legal system. Rules of adjudication confer powers to determine how rules are applied (e.g. whether they have been breached). The rule or rules of recognition are axiomatic. It (or they) determine whether something is, or is not, part of the legal system. Rules of recognition provide criteria which tell us whether something is law (e.g. is it is found in an Act of the Federal Parliament?).[William, 1980, p105 et seq], [Wacks, 1999, p59 et seq], [Leiboff and Thomas, 2004, p168 et seq]

### 2.1.2.3   American Legal Realism

The work of American legal realists approached law as an exercise in empirical pre-diction of what judges will do in legal disputes. Thus Holmes stated: *"The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law."* The concern is to dump theory and focus on empirical investigation of law in its practical social setting (i.e. the court system).[Wacks, 1999, pp138-9] The courts rather than legislators are at the centre. Further, they also emphasise facts as well as rules, as sources of unpredictability in court decisions.[Freeman, 2001, pp799 et seq] Frank identifies two types of realism: "rule-skepticism" (uncertainty resides in the rules themselves) and "fact skepticism" (the facts of cases make them uncertain). Thus, to treat the law solely as a set of rules is profoundly misleading.[Freeman, 2001, pp803-804] In Llewellyn's approach to realism, law is a description of what law does (its functions) within an institutional context, for example its "doing something about disputes".[Wacks, 1999, pp141-2]

### 2.1.2.4   Critical Legal Theory

Critical legal theory is concerned to expose the social "realities" of law in another sense. It echoes and departs from a perspective identified by Marx and Engels: *"Your jurisprudence is but the will of your class made into a law for all."*[Cotterrell, 2003, p209] Although having a number of other dimensions, a central theme of critical legal studies is a conceptualisation of law as an instrument validating and supporting unjust structures of power. Critical legal scholars are concerned to demonstrate and critique the law's role in maintaining unequal power relationships in society.[Letwin, 2005, 260-261] Thus for example, contract law attempts "to conceal what is going on ... like other images constituted by capitalism".[Wacks, 1999, p219] Gabel points out that legal concepts are ways of playing with concepts that are given the appearance of "things", or actual "existence" (i.e. reification). Believing that we are "actually living in a world of rights-holders, legal subjects and formal equality" legitimates the status quo.[Freeman, 2001, p1048] Such concepts are part of a broader emphasis on the socially constructed and relative nature of truth drawing on social theory.[Freeman, 2001, p1051] While they are now seen as passé, they were the matrix in which critical race theory and critical feminist jurisprudence developed.[Freeman, 2001, p1055] In a sense, critical legal theorists sought to overturn the separation between socio-political and legal that Austin and his followers believed to be essential to understanding the law.

**2.1.2.5   Feminist Legal Theory**

Discussion of feminist contributions to legal theory typically trace a trajectory of evolution from liberal feminism (which focussed on equality of treatment and rights within the prevailing legal system) to radical feminism (and beyond) which re-conceptualised the legal system as essentially patriarchal in its characteristics and therefore oppressive of women and society as a whole.[Cotterrell, 2003, p215],[Barnett, 1998, pp 57 et seq] A claim sometimes made for the law is that it is neutral as between social actors. Feminism challenges the claimed neutrality of the law.[Penner et al., 2002, p779] Radical feminists view the law and legal theory as essentially male. Thus Olsen argues that historically certain dualisms in how the world is interpreted: such as rational/irrational, active/passive, thought/feeling, culture/nature, objective/subjective are sexualised (male vs. female), have embedded a hierarchy in which the masculine is privileged and the law lines up with the masculine side of the dichotomies.[Penner et al., 2002, p 803] While assessment of society as essentially patriarchal is at the heart of feminist analyses, those analyses are diverse and reflect different movements within feminism.[Freeman, 2001, p1124],[Barnett, 1998, pp 57 et seq] An interesting direction in feminist thought is in its analysis of power. Some feminists broaden concepts of power, contrasting traditional patriarchal concepts which involve "power over" others, with concepts such as "power to" (the inherent capacity of a person to engage in personal and societal transformation) and "power with" (the power emerging from the collaboration of social actors); or "power within", a subjective psychological self-confidence or self-esteem.[Sisson Runyan, 1994],[Tickner, 1994],[Reingold, 1996],[Blakeley and Bryson, 2007, p146 et seq] The Austinian sovereign-subject dichotomy is very much in the 'power-over' mode.

As far as I am aware, feminist thought has not built a theory of law based on these alternative conceptions of power, although it might have sought to do so. It is useful here to refer also to Michael Karlberg's analysis of power. Given the centrality of power to law - better understanding the nature of power offers insights into the nature of law. Karlberg approaches power through the lens of cultural theory. He maps power in two dimensions. Firstly a dimension querying how law is distributed between human beings. All humans have power. In this sense, power is the "power to", or "power within" each human person. However within a social and cultural context such power is differentially distributed. Law makers, or those who most strongly influence law makers, for example, have greater power than the subjects of law. A second dimension of power is how human beings choose to use the power they possess. Power may be used on a spectrum from adversarial to cooperative.[Karlberg, 2004] (See Figure 2.1 adapted from [Karlberg, 2004]) This two dimensional account

of power allows us to see power (and by implication law) mapped against these cultural dimensions. Thus, the red dotted box represents conceptions of law from a "power over" view of power. Both critical perspectives and positivist perspectives fall within this area of Karlberg's mapping. The remain three quadrants provide spaces in which feminist conceptions of the nature of law might be further explored. What does it imply for the nature of law if it is not conceived primarily in "power over" terms, as virtually all schools in all contexts, do? Feminist theories of power are further examined in Section 2.1.8, in the context of design approaches to law.
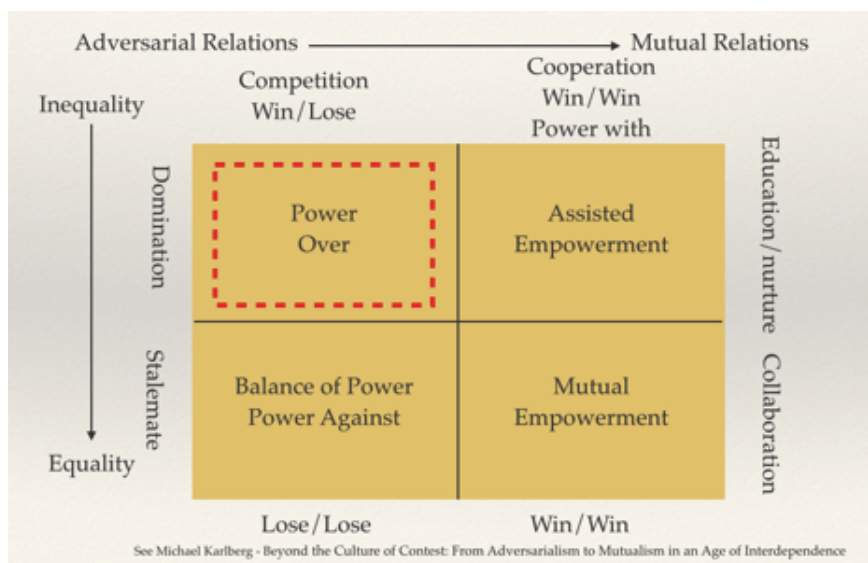


Figure 2.1: The Nature of Power

### 2.1.2.6 Jurisprudence - Reprise

This brief review is an admittedly limited investigation of an enormous field. Yet some themes emerge which are of interest in the context of the examination which follows. With the exception of realist analyses, all theories essentially conceive of law as a set of rules, whether that set is viewed favourably or critically. Thus, the command theory of law unashamedly places commands (i.e. rules) issued by the sovereign at the centre of its conception of law. Critical legal theorists, although insisting that law are rules imposed on the weak by the powerful, still have rule at the centre of their concerns. Natural law in its own way conceives of laws as rules - in this case, rules whose existence can be derived by reason. Feminist legal theory, in its critical dimension, has similar implications: law is patriarchal power imposed by rules.

The American realist approach is essentially a theory of law for legal practitioners. Its vision of those who make the law and use the law is confined largely to a court setting, and to judges, lawyers, and perhaps incidentally their clients, in that setting. The limited social context of this theory of law does not reflect the law's permeation of diverse social spaces, from international relations to the family home. It does not even reflect the realities of legal practice, much of which occurs outside the court system. It is perhaps useful as a tool of professional practice, but we might query how useful it may prove outside that context. Finally, as a theoretical frame, it is anti-access, as non-lawyers are largely excluded from its frame. It does not enable us to ask, indeed does not consider it relevant to ask, how access to law for non-lawyers can be enhanced. From a critical viewpoint, it lends structural support to the power of a professional elite (the legal profession) as against others who either need or desire to use the legal system. The issue of access to law is further discussed in Section 2.2.2.

[Curtotti et al., 2015a],[16] among other things, considers the multidisciplinary insights that might be drawn as to the nature of law from computational, business and design perspectives. It introduces the question of what law may mean to us when one begins to view law from these perspectives. Legal theory, even with its significant diversity, is apt to appear narrow in its investigations when multidisciplinary lenses are applied. From such multidisciplinary perspectives it is clear that law is many things, or that it is a complex phenomenon which is only amenable to simple definition at the cost of over-simplification. The "rules" of law are, for example, a communicative process, and can be studied in that light from a linguistic viewpoint and as a socio-cultural process of communication. From an informatics viewpoint, law is data and information, and can be represented and manipulated as such. From a mathematical perspective, law has mathematical characteristics, such as network characteristics. From the viewpoint of designers, laws are designed artefacts, albeit abstract rather than physical artefacts, and focus is on questions of quality of design. Considering law from these new perspectives enables new questions to be explored and fosters the use of law in novel ways. While, in principle, such investigations could have been carried out in the pre-computer and pre-internet days, the availability of these technologies makes feasible explorations in some of these domains far easier to carry out than would have been the case in the past. The availability and application of computational technologies facilitates the imagination of law in new ways. The following sections discuss some of these dimensions in light of the published works.

They also prompt reflections along a line of approach described as "law as ...".

---

[16]*Interdisciplinary Cooperation in Legal Design and Communication* (Appendix A.3)

This approach does not look at law through use of 'and' (law and field X), but rather looks at law as metaphor. The "law and ..." approach implicitly subordinates law to other scholarly methods and disciplines and makes law only accessible through particular schools (e.g. positivism or realism). Lavi describes this approach colourfully:

> *"Law is akin to a patient who has donated her body to science. Doctors from different disciplines assemble around the bed and hover over the body to study it using the most advanced tools at their disposal. The corpse is still warm, but experimentation has already begun."* [Lavi, 2010]

"Law as ..." by contrast uses metaphor to enable open-ended investigation of law through a potentially limitless diversity of identities. Secondly the "law as ..." approach re-frames consideration of law by re-engaging abstract legal theory with the historical material aspects of law: for example theories of master-servant legal relations with the realities of the workplaces from which the abstract principles were derived.[Mertz and Rajah, 2014] The ellipsis of "law as ..." does not lead to grand theoretical conclusions and rather points towards seeing law in new ways.[Tomlins and Comaroff, 2011] Lavi describes an "ontology" of law - an investigation of its being. Although using the connector "as" in a slightly different sense, he discusses law "as" science, policy and culture. Lavi challenges the assumption that "law and" scholarship does not offer insights into the nature of law. By making the ontological shift to "law as ...", the nature of law can be fruitfully investigated.

> *"... we rarely think of "law and ..." scholarship as offering an answer to the fundamental jurisprudential question of what law is. But, as we shall see, it does. ... I offer a tripartite mapping of legal scholarship that corresponds to three different ontologies of law. ... the theoretical legal landscape can be divided into three categories: law as science, law as policy, and law as culture. ... This tripartite mapping of legal theory ... does not exhaust the entire field of scholarly research. ... Limited as the map may be, it is important for our purposes because it demonstrates both the epistemological and modernist presuppositions of contemporary scholarship. Indeed, science, policy, and culture are not simply three provisional headings in an ad hoc map of legal scholarship. They constitute three different answers to the ontological question: "What is modern law?" "Law as science" identifies the way law is with the way science is; similarly, "law as policy" and "law as culture" equate law's being with that of policy and culture, respectively."* [Lavi, 2010]

An example of a "law as ..." approach is provided by Haapio who looks specifically at contracts beyond their traditional construction as "legal instruments" (i.e.

as repositories of legal rights and obligations). Contracts are also "instruments for conducting exchanges", "for planning and communicating". They are also "things and social artefacts". Contracts enable as well as control.[Haapio, 2013, pp27-40]

The investigations carried out below seek to proceed in the spirit of the "law as ..." paradigm. They do not seek to reach reductive, definitive conclusions as to the nature of law, but rather they seek to further illuminate our understanding of law, by thinking of law "as" these various categories. In doing so, the approach calls into question limitations inherent in adherence to particular schools of legal theory, while leaving open the possibility of drawing on each of them.

What follows, in a sense, moves from a theoretical to empirical discussion of law. What do the investigations reported in the published papers listed at the beginning of this thesis suggest as to the nature of law? It is important to note that the investigation here is not primarily in the mode of "thinking" about law. It rather arises from "doing" law in the context of a body of research applying computational technologies to law. Not doing law as lawyer: rather doing law as communication; doing law as software development; doing law as computer science; doing law as researcher; doing law as design. To take Lavi's medical dramatisation as departure point, law gets up off her death bed, looks in the mirror and discovers she is not what she had been told she was. Her identity, she finds, is hers to remake.

### 2.1.3 Law as Language and Communication

> *"When people talk to each other, they never say what they mean. ... They say something else and you're expected to just know what they mean."*

> Alan Turing, fictionalised statement in *The Imitation Game*, 2014.

In [Curtotti and McCreath, 2011],[17] contracts are approached as a body of language. A corpus (body) of language is compiled and studied for its linguistic characteristics. The characteristics can be compared with other forms of English, for instance finding that prepositional use is far higher in contract language than in general English.[Curtotti and McCreath, 2011] Such linguistic characteristics can be leveraged (as is done in [Curtotti and McCreath, 2010]), to automatically classify text in a contract into functional categories, such as headings, execution clauses or definitions. Having automatically extracted definitions, the complex inter-relationships between definitions can be represented in various ways assisting the task of both reader and drafter, as is reported in [Curtotti et al., 2013].[18] This direction introduces concepts of visual communication or information visualization, an alternative language

---

[17]Appendix A.7
[18]Appendix A.5

with which meaning can be conveyed. The visual communication of law is further explored in [Passera et al., 2014] and [Curtotti et al., 2015a],[19] which discuss multidisciplinary work automating the visualization of contract clauses dealing with contract termination or expiry, payment structures, or liquidated damages. Enhancing the visualization of legislation online is discussed in [Curtotti and McCreath, 2012].[20] The content of legal rules, as manifestations of language are further addressed in [Curtotti et al., 2015c][21] and [Curtotti et al., 2015b],[22] explores the readability of law and how it may be improved through application of computational technologies.

At the broadest level, these papers address enhancing the communication of law through application of computational techniques. However these explorations contain assumptions as to the nature of law. Firstly, legislation, regulations and contract provisions have been taken as the subject matter of study. This fits comfortably within the focus of most of the legal philosophies discussed above, with the exception of American realism and its "rule skepticism". Implicitly the "rules" which express law, or to put it another way, the written language regulating human interactions ("regulatory language"), is considered central to law. Law is a language. Law is a process of communication, and the efficacy of that communication is studied and problematised. While useful to the particular projects involved, they do not present a complete account of the law, by any means. This may be illustrated by two 'early' examples of law making. One from colonial Tasmania in the 19th century (the misnamed Governor Davey's Proclamation); and another, being the earliest surviving example of English law, indeed the earliest surviving example of English as a written language: the laws of Aethelberht.

In the early nineteenth century, English law was introduced to Tasmania. In what was deliberate deception or self-delusion or both, Governor Davey's Proclamation sought to communicate to the indigenous population, principles of equality, the rule of law and punishment for murder, using images instead of words. (See Figure 2.2) Beyond the assurance of 'fairness', the communication presumes the existence and primacy of a newly introduced legal and cultural order. The order itself was however highly contested given its colonial genesis. The idea of such a communication was hit upon by the Surveyor-General based on emulating bark paintings then in use by indigenous Tasmanians. Manderson underlines that the indigenous population experienced none of the promised benefits of the proclamation.[Manderson, 2012]

Here communication is attempted through image, an already accepted technology of communication. It provides a striking, if sadly deceitful, example of commu-

---

[19]Appendix A.3
[20]Appendix A.6
[21]Appendix A.1
[22]Appendix A.2

Figure 2.2: Governor Davey's Proclamation

nication of legal principles and rules using images. More powerfully, the example points to what critical and feminist theorists suggest about law: that it may represent the imposition of the will of the powerful on the powerless. Indigenous conceptions of law and its application in the circumstances are likely to have lead to quite different conclusions. Further, the addressees of the communication are hardly likely to have shared the implicit assumptions of the creators of the images. These assumptions give primacy to European conceptions of civilisation and human relations and to the existence of a legal order which was itself contested. The "rules of recognition", to draw on Hart's schema, are entirely different for the two communities. This example of law making can hardly be understood outside its social and historical context.

A similar process of introducing new legal concepts occurred thirteen centuries earlier in Anglo-Saxon England when the Laws of Aethelberht were introduced. Again the genesis involved a meeting of cultures. Written in the 7th century, they were adopted shortly after the arrival of the first Roman-Christian mission. The laws

are dominated by the theme of compensation for injury or wrong, suggesting a society beset by violence. The complex social implications of law are in this case raised by a conscious prioritisation of Church and King. The laws were used to elevate the social status of these institutions by instituting higher compensatory payments for injuries to king, church and religious office holder.[Attenborough, 1922] The circumstances of the creation of Aethelberht's laws are speculative, given the paucity of contemporaneous sources. Why writing was resorted to, when oral transmission was the norm, is debated. The influence of the Christian mission is offered as an explanation. "Emulation of the Romans", is suggested by Bede. Assertion of a super-kingly status through the creation of a written record is offered as another.[Oliver, 2002] The use of the written word is far more than merely an exercise in language for the purpose of instituting rules.

These two examples underline that law is created within a socio-cultural context and to the other personas of law must be added that of communication. Law is an act of communication between human beings whose relationships, purposes and interactions are complex. The examples also highlight that the technology of communication makes a difference to those involved in the communicative act. Communication can be embodied in sound, symbol or image. They highlight further that communication can be multilayered, simultaneously designed to convey different meanings or to pursue different purposes with respect to different audiences. As seen above, either to limit conceptions of law to communication or to leave out the perspective of communication, is to leave out dimensions which are necessary to a reasonably complete account.

[Allen and Engholm, 1978] provide an interesting intersection of considerations of law as language and the use of image to represent legal meaning. As they state, "[i]t is hard to exaggerate the importance of language as law". They identify a number of ways in which communication (with more acute consequences in the legal case) can be problematic. Uncertainty, ambiguity, impression and incompleteness are all examples of problematic communication. Such problems can arise in semantics (word meaning) or in syntax (sentence structure). They propose the use of "normalized drafting" to address syntactic ambiguity in particular in the law. Thus an example is given of legislative ambiguity where a statute might say a person shall not do "X and Y". Does this mean a person shall not do X and a person shall not do Y; or does it mean a person shall not do (X and Y)? They propose the use of diagrams (flow charts) and careful paragraphing to make sure that this kind of ambiguity does not inadvertently arise. They demonstrate how readily flow charts can be used to clarify the meaning of layered conditionals and conjunctions in legal provisions.[Allen and Engholm, 1978] Notably while their proposals represent address syntactic problems,

they do not address the inherent fuzziness of word meanings themselves. Here "X" and "Y" may be far from clear, even after syntactic ambiguity is removed.
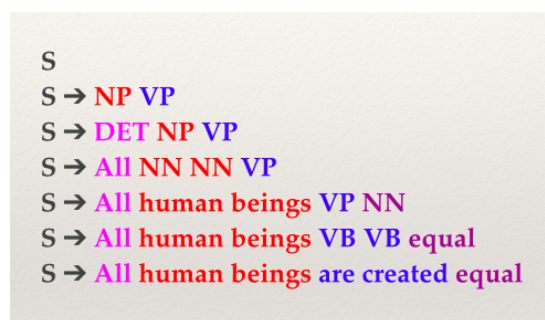
It may also be observed that law as language is different from language in general. In this respect, law constitutes a sub-language: i.e. one of many "registers" or "genres" that exist in any language. Such registers may be applicable to certain situations (family vs. public spaces), particular professional groupings (lawyers, doctors, engineers), particular geographic or demographic groupings. Within law itself there are a variety of registers. The language of contracts, legislation, the court room, the police interrogation and legal academic discourse are all distinct, yet all within the scope of law as language. Further, language varies from jurisdiction to jurisdiction, with, for example legislative styles being more general in civil law countries as compared to more particular in common law countries.[Jackson, 1995, p89-96] Even within common law countries (for example Australia vs. the United States) legislative registers may be notably different. Jackson suggests that the professional registers (plural) of the law are internally problematic for it.

> *"On the one hand [law] claims to be a unified system, a standard accessible to all: law is expressed, for the most part, through (a particular register) of natural language. ... even non-lawyers have access to it, as indeed the ideology of law requires. On the other hand, the law clearly possess a culture of its own. Even **within** the legal community, there exist many different occupational groups, each with its own version of legal language."[Jackson, 1995, p97]*

Questions of power are also associated with language. This is most obvious when the entity generating the language is doing so from a position of power. More complex expressions of power relations are found within social situations. For example male speech being found to express more "power" than female speech in some studies in a courtroom context.[Jackson, 1995, p99-110]

Computer scientists are also concerned with languages - although most often as languages usable by machines; or languages usable in human-machine communication. Considering these computational perspectives provides an interesting view of language and in that context of the law as language. Computer scientists define formal languages as a subset of languages. A formal language will be built from a finite set of symbols. The language is a set of sentences (combinations of symbols) that can be validly formed using the symbols. A language has a grammar which imposes structure on the symbols (the rules which specify the allowable sentences in the language). The representation of a grammar can be undertaken by the use of symbols of a metalanguage for the description of the language. These symbols (non-terminal symbols) can be used to generate or analyse valid sentences in the language.

A chain of substitutions can be undertaken by which an initial non-terminal symbol S is replaced by other symbols using substitutions permitted by the grammar, until all non-terminal symbols of the metalanguage have been removed. (As for example in Figure 2.3) Thus a formal language consists of that subset of all possible sentences that are valid in the given grammar. [Grassmann and Tremblay, 1996, Chapter 10] [Russell and Norvig, 2003, p790-792] Figure 2.3 illustrates this process with a subset of English. Non-terminal symbols such as noun phrases and verb phrases are replaced with other symbols until only terminal symbols (in this case words) are left. This kind of framework is essentially what makes possible the creation of software programming languages which enable humans to "communicate" with machines. In programming languages human understandable symbols serve as non-terminal symbols of a metalanguage that are substituted in the computer ultimately to 1's and 0's which are the terminal symbols of computation.

S
S → NP VP
S → DET NP VP
S → All NN NN VP
S → All human beings VP NN
S → All human beings VB VB equal
S → All human beings are created equal

Figure 2.3: Symbols and substitutions - a computational approach to language

It is possible to also extend such formal languages to represent the semantics, or the meaning of symbols. In this case symbols are related by a logical grammar rather than syntactic grammar, but the process of symbol substitution remains essentially the same. Thus a representation such as **loves (john, layla)**, can be used to represent the semantic subject-relation-object relationship involved. This of course points to an essential characteristic of language as a carrier of meaning between agents involved in the act of communication.[Russell and Norvig, 2003, p794,811-815]

This seems distant from conceptions of human language let alone law as language. However the extraction of syntactic relations using the more elaborate methods on natural language processing, which build on formal languages, can contribute to an understanding of the effectiveness of legal communication, as for example undertaken in [Curtotti et al., 2015b][23] in studying the readability of law. In respect of logical relations, such structures form the foundation of expert systems able to treat

---

[23]Appendix A.1

law as computation. Computation with law of this kind is very much in a positivist mode. Rules, fact and the application of rules to facts are all certainties capable of logical analysis. Law as computation is further discussed in Section 2.1.7.

As we have seen above, human language does not share the certainties of formal language. Words, the primary symbols of human language, are typically indistinct and often multivalent in their meaning. Moreover, as words are combined into larger units of language, possible syntactic ambiguities and meanings expand exponentially. Parts of speech attributable to a word may be uncertain. In the Brown Corpus of English up to 40% of words used in the corpus have ambiguous tagging.[Jurafsky and Martin, 2009, 133] Use of probabilistic methods can however predict tags with a 97% accuracy. This is similar to the agreement among human annotators.[Jurafsky and Martin, 2009, 155]. This is, however, only the first level of ambiguity. Parsing English into grammatical phrases introduces potentially exponential possible parses as sentences grow in size. Again probabilistic techniques are used to address the problem, achieving around 90% precision and 90% recall on the Wall Street Journal corpus.[Jurafsky and Martin, 2009, 480] This is without attending to ambiguities of meaning. For example the word "bass" may be a fish or a musical instrument. Many words have such multiple meanings and uses.[Jurafsky and Martin, 2009, p638 et seq] Further as we saw above, communication involves social complexities that are not necessarily evident in the communicated message itself.

What do such computational results tell us about the nature of law as a body of language? Most obviously, such results call into question scientism suggested by some theories of law. The meaning of language is not specified with certainty; at best we can only infer probable meaning. However, does legal language perform better than general language? Persuasive evidence that it does is lacking. Further there is persuasive evidence in the opposite direction. A review of the research literature reported in [Curtotti and McCreath, 2013][24] and [Curtotti et al., 2015c][25] establishes that readability of law is low for most audiences (in the difficult to incomprehensible range). This is despite the desire of writers of law to achieve precision in meaning. Precision comes at the cost of comprehensibility. As readability studies make clear, the readability of text only has meaning in relation to a specific audience.[DuBay, 2004] Further the purpose of language is conveyance of meaning between communicating agents. If an audience finds law incomprehensible, little meaning at all is conveyed, and ambiguity of meaning becomes a secondary consideration.

Of course, to approach the language of the law in this way, is in some respects to impose on its creators intentionality that may have been low in their priorities (if

---

[24] Appendix A.4
[25] Appendix A.2

present at all). Writing in 1969, Crysal and Davey regarded legal writing as largely archival in its intent - to create records for use by legal experts.[Jackson, 1995, 112]

> *"To speak of legal language as communicating meaning is in itself rather mis-leading. Of all the uses of language it is perhaps the least communicative, in that it is designed not so much to enlighten language-users at large as to allow one expert to register information for scrutiny by another."* [Crystal and Davy, 1969] cited in [Jackson, 1995, 112]

Such approaches to the law are echoed in the priorities sometimes expressed by legislative drafters, for example in the following observation by a legislative drafter which is primarily concerned with the process of law making and law interpretation by public agents concerned with the law:

> *"The drafter has two main objectives. The Bill must be drafted to pass and it must work as intended when it becomes an Act. If a Bill is to pass, it has to be in a form acceptable to a majority of the members of each House of Parliament, and if an Act is to be effective it has to be given the meaning intended for it by the Government when it is construed by the highest court of appeal ... There are subsidiary objectives that also need to be considered ... Every drafter is ultimately seeking to produce a provision that is clear enough for even opposing parties to understand it in the intended sense without unnecessary litigation and its attendant cost."* (Steven Laws cited in [Stefanou and Xanthaki, 2013, pp 24-25].)

As we saw above, placing citizens at the end of drafting priorities is problematic. The more so, when notions of rule of law, are expanded to a consideration of the imperatives of democracy.

> *"It is strange that free societies should ... arrive at a situation where their members are governed from cradle to grave by texts they cannot comprehend."* [Bennion, 1983, p8]

Recent years have seen a seismic shift in the communicative context of law. In the 19th century the cost of the statute book in the United Kingdom was equivalent to the annual wage of the average worker. [Bentham, 1843] In the past, it didn't matter so much if the general public couldn't read the law. A member of the public rarely attempted to do so. In the late 20th and early 21st century, the Free Access to Law Movement, as well as official and commercial publishers have increasingly ensured that the law is available to everyone through the world wide web. Comments by

official publishers, for instance the UK Archives Office pursuing the "Good Law" initiative, and research results of the joint research project of the Australian National University Research School of Computer Science and the Cornell University Law School Legal Information Institute studying the demographics of the reading of legislation online, suggest that lawyers are now a minority of those who read (or at least attempt to read) legislation.[Curtotti et al., 2015c][26]

While the 'people' may be increasingly reading the law, it appears, ironically that their law makers are less likely to be doing so. At least in the United States, voting for a bill and reading it do not appear to have any necessary connection. [Jones, 2013] This has obvious democratic implications. Beyond this, it has implications for the volume of law as well as its quality. While noting that the evidence is that elected representatives rarely read legislation in the U.S. Congress, Jones argues that law makers ought not be required to read the law as this allows them to specialise. [Jones, 2013] Apart from the issues above, taking elected representatives out of the reading public for the laws they make creates a process that drifts further and further from conceptions of democratic governance.[Jones, 2013] Bennion's irony is incomplete. It might better read:

> *"It is strange that members of free societies are governed from cradle to grave by laws which they cannot comprehend. Stranger still, elected law makers in such societies rarely take the trouble of reading those laws before enacting them."*

Whatever the reading habits of lawmakers, a strengthening consensus has in various ways problematised and addressed the communicative problems of legal language. This direction is captured in the work of what is called the 'plain language' or 'plain English' movement. The movement has been influential and legal language is not what it was. Concerns about legal communication stretch back centuries.[pp 124 et seq, pp 133 et seq][Mellinkoff, 1963] The nineteenth century saw successful reform of legal language that transformed legislative writing and established professional offices with responsibility for it.[Bowers, 1980], [Bentham, 1843, pp 250-251],[Evans and Jack, 1984], [Renton, 1975] The twentieth century saw increasing interest in plain language in the writing of law. Works such as *"The Language of the Law"* by Mellinkoff drew attention to the issue.[Mellinkoff, 1963] Indeed laws have been passed mandating plain language in government and legislative drafting - for example the Plain Language Act 2010 (U.S) and an executive order issued by President Barack Obama in 2011 requiring regulations to be written in plain language. (See further discussion in [Curtotti et al., 2015c][27])

---

[26]Appendix A.1
[27]Appendix A.2

Examination of law as language is further explored in Sections 2.2.3 and 2.2.4, which respectively address the readability of law and visual communication of law. What is clear from this brief review is that rich insights as to the nature of law can be drawn from a consideration of it "as" language.

### 2.1.4   Law as Document

*"What is law? Is it what is on the books, or what is actually enacted and obeyed in a society? Or is law what must be enacted and obeyed, whether or not it is on the books, if things are to go right?"*

Bernard Schlink, *The Reader*, p89

In [Curtotti and McCreath, 2012][28] a review is provided of the publication of law online. The publication of law as document is nothing new. While the material form has changed, the documentary character of a modern statutory document has little to distinguish it from the tablets of stone and clay which are its ancestors. In the modern context, the concept of document survives as virtual document in the online environment. Thus, Australia's official federal legislative site[29] provides a "current" version of a statute, and a list of previous historical versions, each of which can be accessed as a virtual document in pdf or word format. A similar approach is taken by the Australian Capital Territory, which publishes compilations of its laws at particular points in time. This approach simply continues traditional printing practices which were a necessary consequence of paper based printing. While far from universal, with other jurisdictions and publishers exploring alternative digital publication models, they illustrate the power of the documentary paradigm in the legislative space. Another example of the documentary paradigm is offered by a review of the presentation of Australian legislation undertaken in the mid-1990s. Among seriously held concerns were concerns as to the publication costs if fully marked up versions of amending provisions were prepared and concerns that readers would be distracted by underlining of defined terms.[Curtotti and McCreath, 2012] Surden provides another example of law as document. He notes that a barrier to the automation of contracts has been that contracts (conceived as written instruments) have been regarded as beyond the capacity of computers to analyse and automate.[Surden, 2012]

By contrast, the documentary paradigm is largely excluded from the education of law students when addressing the topic of contract law. Rather than contract law being concerned with concrete actual examples of contracts (the private law analogue

---

[28] Appendix A.6
[29] [Australian] Federal Register of Legislation http://legislation.gov.au

of legislative documents), contract law primarily concerns itself with abstract legal forms of which contracts are comprised in legal theory. Law students will rarely (if ever) see a contract, much less be asked to draft one. Drafting is an afterthought (in the Australian context), taught in a brief professional skills course meant to transition a law student from the academic context of law school to the realities of legal practice. This culture contrasts with the education of computer science students who are immersed in the creation of software artefacts from the outset of their education.[Curtotti et al., 2015a] It is interesting to observe that legal education may represent an outlier in comparison to professional education in many other fields: e.g. medical, allied health, engineering, science, where education and professional practice are much more closely related. In contrast to contracts, legislation is read in law school, and the contrast points to the nature of law as data: law is studied as case law and legislation in law schools because that "data" is readily available to law schools. Until recent years there have been no readily accessible data repositories of private contracts.[30] In Section 2.1.5, law is explored as a form of data.

These examples illustrate how 'law as document' is both central and marginal to concepts of law. In the first case, the implicit documentary model limits the potential of law in online environments. In the second, a lack of appreciation of the documentary dimensions of law closes off potential directions in legal education. This example encourages us to view concepts of law, not as truths to be established in opposition to competing views, but rather as conceptual tools that enable more effective application of law.

### 2.1.5   Law as Data, Information and Knowledge

*"We will show them Our Signs in the universe, and in their own selves."*

Qur'an, Sura 41:53

Thinking of law as data introduces new dimensions not explored above. It is an essential step in applying computational technology to law. This is clearly illustrated in the contrasting manner in which the UK has approached the publication of its laws as compared to many other jurisdictions. The UK Archives Office, which is responsible for publication of the statute book online, did not conceive itself as placing virtual copies of documents online (although its system is able to produce such virtual documents). Rather it began by viewing the statute book as a repository of data.

---

[30]An example of such a repository is the U.S. Securities and Exchange Commission filing pages which includes contracts material to the economic position of companies. *Researching Public Companies through EDGAR: a Guide for Investors* http://www.sec.gov/investor/pubs/edgarguide.htm accessed on 4 December 2015.

It organised that data by creating an xml representation of the entire statute code. This xml representation contains more than the text of the laws. It contains metadata relating to each provision in the statute book. For example, where applicable, the functional character of particular text is included as data in this xml representation. Returning to the formal language used by computer scientists, this is essentially what is done in an xml representation. The 'terminal nodes' - the actual language of the statutes, is enriched with a set of non-terminal nodes describing the data. The data is made available to developers who may wish to use that data in a variety of ways. This approach enables UK legislation to be navigated at section level and to be enriched with information about a provision, such as when it was amended, what versions were in effect at different points in time, defined terms and annotations relevant to a provision. When law is thought of as data it can be provided as input to a computational process. (See [Curtotti and McCreath, 2012])[31] The conceptualisation of law as data opens the possibility of a wide diversity of uses, of which the UK legislative site is only one of the most leading examples. In [Curtotti et al., 2013],[32] natural language in contracts is converted to enriched data which is used as input to visualizations of definition networks within contracts. Each definition is treated as a data point and node within a network of nodes.

Surden discusses another important application of law as data. He reviews the rise of "data-oriented" contracting. Surden defines this as contracting where the contracting parties have decided to represent some subset of the contractual terms as data with the intent that the contract be partially machine readable. This can be achieved by created structured data (for example xml representations of data in a contract). While computers find it difficult to attain 100% accuracy with natural language; the same is not true in respect of structured data. Agreement among human participants as to a form of data representation potentially makes that data available to computational technologies. Among the benefits suggested by Surden for data oriented (and computable) contracting are reduced transaction costs, improved monitoring, new analytic applications and computer to computer transactions.[Surden, 2012]

Data is related in meaning to the concepts of "information" and "knowledge". Boisot and Canals, writing as part of the academic discourse of economics, make the point that our understanding of these three terms ("data", "information", "knowledge") is "vague". They cite the example of cryptography to provide an intuition into the difference between data and information. An encrypted message is undeniably data. It also contains information. That information is inaccessible without a key

---

[31]Appendix A.6
[32]Appendix A.5

to decode the information in the message. The difference between information and knowledge on the other hand they explain by positing a sentence: "The cat is tired". What this sentence means depends entirely on the context of the message. "The cat", for example, may carry a meaning that can only be understood with knowledge of its surrounding circumstances. They give the example of a Mafia boss known as "the Cat": "To understand the sentence is not necessarily to understand the message." They define information as "an extraction from data that, by modifying the relevant probability distributions, has a capacity to perform useful work on an agent's knowledge base." This agent based view, they further elaborate by understanding the distinction between data and information in terms of processing by an agent. Data being selected by an agent's sensory functions, and information being selected by its conceptual filter. Knowledge is processed information incorporated into the agent's conceptual models and on which the agent is able to act.[Boisot and Canals, 2004]

Zins, looking at the conceptual distinctions between data, information and knowledge from the viewpoint of Information Science, similarly begins his treatment cautiously.

*"The academic and professional IS literature supports diversified meanings for each concept. Evidently, the three key concepts are interrelated, but the nature of the relations among them is debatable, as well as their meanings."* [Zins, 2007]

He notes that much of the literature proposes a sequential ordering of data –> information –> knowledge. One view, he notes, is that knowledge only exists in the mind of a person. Another view suggests that information and knowledge are largely synonymous. Zins proceeds to analyse the views of 45 leading scholars participating in the Critical Delphi Study on the meaning of data-information-knowledge (D-I-K). His summary of these views emphasises a number of models depending on whether data, information and knowledge are viewed as objective or subjective entities or both. The most common view was that data and information are objective while knowledge is subjective. Other issues canvassed in scholars' views included whether D-I-K related exclusively to human agents or whether it extended beyond humans (non-human agents, the natural world). Among views of information, were those which emphasised its connection with communication.[Zins, 2007] The communication view reflects the origins of information science in Shannon's definition of information in relation to its communication through a constrained channel. ([Shannon, 1948] cited in [Boisot and Canals, 2004]) Figure 2.4 summarises Zin's synthesis of definitions of D-I-K depending on whether a subjective or objective view is taken of them. The summary assists in seeing how these terms may be being used in the lit-

erature and how they might be related to law as data-information-knowledge.[Zins, 2007]

| | Subjective | Objective |
|---|---|---|
| **Data** | sensory stimuli which are perceived | a set of symbols representing stimuli |
| **Information** | meaning ascribed to sensory stimuli e.g. sound == music | a set of symbols representing knowledge |
| **Knowledge** | belief/thought within an individual | a set of signs or symbols representing meaning ascribed by an individual |
| *Simplified D-I-K definitional schema as described by Zins* | | |

Figure 2.4: Simplified Zin's Schema: Data - Information - Knowledge - Subjective and Objective Frames

Chen et al. examine the meaning of D-I-K, in the context of information visualization. Similarly they begin by noting that "there are many competing definitions" of these terms, across a variety of disciplines. Even within the scope of visualization, the terms "remain ambiguous". They note that a hierarchy of concepts data - information - knowledge - wisdom is a popular framework. Like Zins, they distinguish between the subjective and objective meanings of D-I-K, although in their version the distinction merely implies a particular location for D-I-K ("cognitive space"), as opposed to a location in a computer ("computational space"). Rather than rigorously defining each concept, they content themselves with noting that data and information are overlapping as are information and knowledge. [Chen et al., 2009]

> *"Since we can read data, grasp information, and acquire knowledge, we must differentiate these terms in the perceptual and cognitive space. Because we can also store data, information, and knowledge in the computer, we must also differentiate them in the computational space."* [Chen et al., 2009]

This framework is then used to describe different visualization processes that may involve representing varying levels of abstraction in a visual representation. For example, a visualization may simply present data. Data may be processed into information within the computer and that information is then visualized. Further processing may enable knowledge to be presented (for example logical representation within the computer may allow knowledge concerning data to be extracted and

represented visually).[Chen et al., 2009]

The brief literature review above can be related to both of the scenarios of the communication of law using writing and the visualization of law.

Boisot and Canal's observations fit well with the scenario of a reader attempting to read law. Legal language is to some extent "encrypted" to the lay user. Information is hidden in what may be incomprehensible language. As pertinently, even when information in a sentence is "understood", knowledge of its intended meaning may not be achieved; as for example in the case where an Act uses a defined term in a way that is contrary to its expected meaning. The distinction between data, information and knowledge is relevant to the communication of law. In this respect, work surveying whether users find a legal sentence to be readable, as reported in [Curtotti et al., 2015c][33] and [Curtotti et al., 2015b],[34] only gets at whether the reader is receiving "information" in the sense used by Boisot and Canals. It does not provide insight as to how or whether "knowledge" is communicated. In other words, testing understanding using cloze tests (essentially a form of comprehension test) or subjective claims of understanding (Likert or semantic tests), cannot tell us about a cognitive (subjective) state of knowledge.

It is also interesting to consider the visualizations of definitions in contracts undertaken in [Curtotti et al., 2013].[35] The visualizations are built on data, information and knowledge drawn from both the contracts themselves and the legal domain. Natural language is used as the input data. That data is enhanced into the form of "information" by tagging the data with its functional role within a contractual document. For example, the first step in visualizing definitions is to capture information of what parts of the data are "definitions". Knowledge of how defined terms are written within the Australian jurisdiction provides a basis for reliable extraction. Provision of a navigation system allowing a reader to navigate defined terms within defined terms is based on knowledge of how lawyers may seek to read definitions and provides knowledge more readily on the meaning of a defined term. The "definition cloud" representation of definitions within a contract represents information concerning which definitions contain hidden meaning and are heavily used in the contract. It also provides an overall image of the content of a contract. (See [Curtotti et al., 2013] for visualizations concerned.)

Adopting the D-I-K hierarchy also points to a consideration of the representation of "the law" as knowledge. What would such a representation consist of? As Faria et al describe, ontologies are a way of representing and mapping a knowledge do-

---

[33]Appendix A.2
[34]Appendix A.1
[35]Appendix A.5

main. They are a way of representing *knowledge*, within the computational space. The representation of a knowledge domain using ontologies can support computational decision making, such as expert systems. An ontology may be formally defined as a tuple as follows.[Faria et al., 2013]

$$O = (C, H, I, R, P, A) \tag{2.1}$$

Where:

C = a set of entities (either a class of entities or an instance of a class of entities) e.g. X is the defendant in a particular case

H = a set of taxonomic relationships in which $c_1$ is a "kind of" $c_2$ e.g. a treaty is a kind of legal instrument

I = a set of instance relationships e.g *"the UN Charter"* is an instance of an *international treaty*

R = a set of non-taxonomic relationships such as: judges *decide* cases

P = a set of properties associated with an entity

A = a set of axioms, rules or constraints in the knowledge system (e.g. a precedent of a higher court has higher weight than the decision of a lower court). (See [Faria et al., 2013])

Santos et al. note that an ontology may be defined at various levels: high-level, domain, task or application. This classification of ontologies is said to promote reuse. For example a domain ontology for law is potentially usable in many computational contexts, while a specific task such as precedent identification will require an ontology which specifies entities related to that task (e.g. the courts and judgements of a specific jurisdiction).[Santos et al., 2013]

A rough parallel in a non-computational domain that provides a greater intuition as to the nature of ontologies, is that they are essentially machine readable thesauri. Their relevance to the question of the nature of law is that they point us towards a synthesis of characteristics rather than a reductive analysis implicit in seeking to extract from legal phenomena the core irreducible minimum definition of law. For example, defining law in the way that is done in Hart's discussion of legal rules is such an analytic approach. Ontologies by their nature seek to synthesise complex phenomena into a relational whole. Building a model or map of all things and relationships in the law is another way of saying: this is law. Everything that might appear within a legal ontology is in some sense "the law". A person *X* who plays the role of witness *W* in case $C_1$ where judge *J* applies law *L* written by parliament *P* or draws on precedent *Z* in case $C_2$, are all in this sense relevant to our understanding of the nature of law. What a business leader does with contracts is law. A politician's

participation in the creation of law, is law. The interactions of lawyer and client in a client interview is law. Treaty making by diplomats is law. Hart's three-fold categorisation of legal rules can be interpreted in this context of small subset (a sub-ontology) of a much larger ontology of "the law". It is valuable in its illumination of the kinds of legal rules that exist within a legal system, but divorced from the broader ontology, is necessarily constrained in its descriptive power.

This insight arises from approaching law as knowledge, which is in turn built on thinking of law as data and information.

### 2.1.6 Law as Network

*"All I'm saying is simply this, that all life is interrelated. And we are caught in an inescapable network of mutuality, tied in a single garment of destiny – whatever affects one directly, affects all indirectly. For some strange reason I can never be what I ought to be until you are what you ought to be, and you can never be what you ought to be until I am what I ought to be. This is the interrelated structure of reality."*

The Reverend Dr. Martin Luther King, *The American Dream*, 1964.

The foregoing discussion explicitly and implicitly raises networks as another frame of reference. Ontologies, just explored, point in the direction of relationships between entities. The law consists not just of things - but of relationships between things. Kelsen's theory of law discussed in Section 2.1.2.2 conceives of norms as punishment norms networked with foundational norms. In Section 2.1.5 reference is made to work on visualization of definition networks in contracts.

In mathematical terms, a network (or a mathematical graph) is an entity consisting of a set of nodes and edges (or connection) between nodes. Formally: $G = \{N, E\}$. Such graphs have mathematical characteristics related to their content and shape. Among such characteristics of graphs are 'centrality, 'degree', 'path length' and 'density' among others. Centrality measures the location of a node within a graph - whether it is at the edge or near the centre of a graph. Degree measures the number of connections to a node. Path length defines the distance between nodes counted either by number of intervening nodes or weight of connections. Density describes the relationship between the number of nodes and the number of connections between nodes.

Figure 2.5 is an example of a mathematical graph. It consists of 8 nodes and 10 edges. Its density is 0.357 calculated as the ratio of edges to total possible edges as follows:
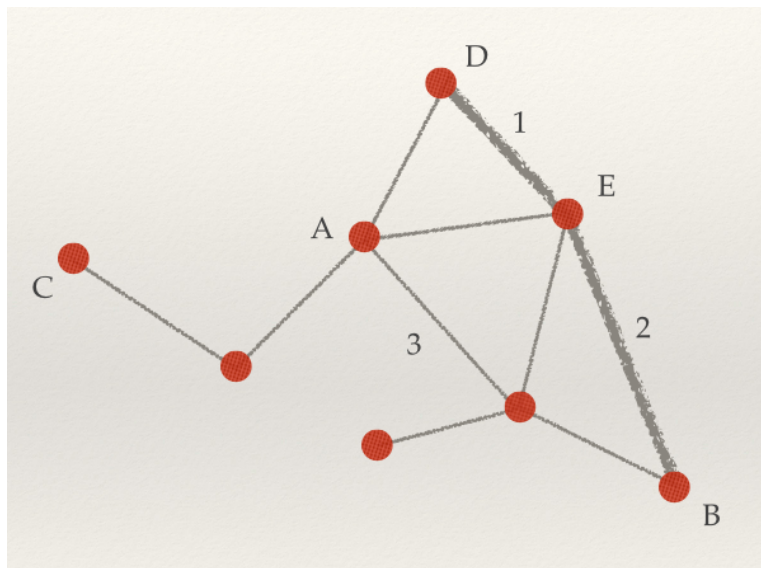
Figure 2.5: Example of a Mathematical Graph

$$Density = \frac{2 * |E|}{|N| * (|N| - 1)}$$

Node A is more central in the graph than node C. Node A has a 'degree' of 4, as it has 4 edges connected to it, whereas node B has a degree of 2. The minimum path length (the shortest distance) between nodes B and C is 4. Nodes A, D and E form a cycle (i.e. a circular path). A graph may have a star like shape (one node connected to all other nodes which are not connected to each other), or more like a circle or line (each node connected to only one other node). Weights may be associated with edges between nodes. Thus in Figure 2.5, edges 1 and 2 are thicker to represent stronger connection than the edge labelled 3. Edges may be undirected (as in Figure 2.5) or may have a direction, forming a directed graph. A graph without cycles is a tree. For example the table of contents of a law forms a tree. (See [Bronstein et al., 2007, pp348 et seq])

By adding information to nodes or edges, graphs may also represent entities in the real world, in which case they are sometimes referred to as networks.[de Nooy et al., 2005, p7] For example, a network may represent relationships between people with each node representing a person. A family tree is such a social network. The grammatical relationships of words in a sentence form a network. Applications such as Facebook which exploit social networks have made the concept of networks better known. Networks are pervasive in reality and law is no different in this respect.

Examples from the law include the relationship between cases in a network of common law precedents and the network of cross-references between sections of a law. The former is for example utilised by RavelLaw to assist the navigation and visualization of U.S. court precedents. The network includes characteristics of nodes (such as their time and originating court) which further enrich the visualization.[36] (See Figure 2.6 which illustrates the citation network connected with Brown v. The Board of Education.)
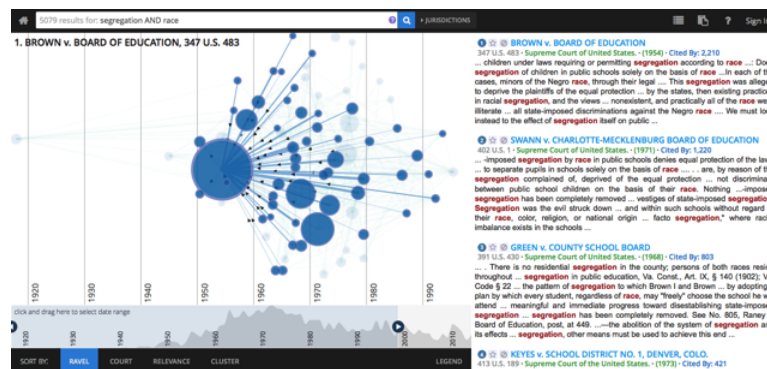


Figure 2.6: RavelLaw Visualization of U.S. Court Citation Network

The AustLII visualization of legislation is based on each section forming a node in a network. Connectors are provided connecting each section with the structure of the law in which it is found, cross-references to other sections, connections to definitions and connections to relevant supplementary materials. Again, exploitation of the network characteristics of law facilitate navigation and of legal content.[37] (See discussion in [Curtotti and McCreath, 2012])[38] Bommarito and Katz carry out a study of the network characteristics of the United States Code. Each section of the Code is represented as a node. Nodes are interconnected by cross-references, and by the formal structure of the U.S. Code. Nodes have internal characteristics based on their linguistic content. They show that over time, the complexity of the U.S. Code is increasing based on the number of nodes, the number of interconnection between nodes and in terms of the information content of each node.[Bommarito and Katz, 2010] In [Curtotti et al., 2013],[39] definition networks in contracts are extracted and visualized. Among the visualizations, is a bimodal representation in which sections of a contract form one type of node and definitions form another. This enables a matrix representation in which shading represents the strength of connection between

---

[36]RavelLaw website https://www.ravellaw.com/
[37]AustLII website http://www.austlii.org.au
[38]Appendix A.6
[39]Appendix A.5

a definition and a section in a contract, providing a heat map as illustrated in Figure 2.7.
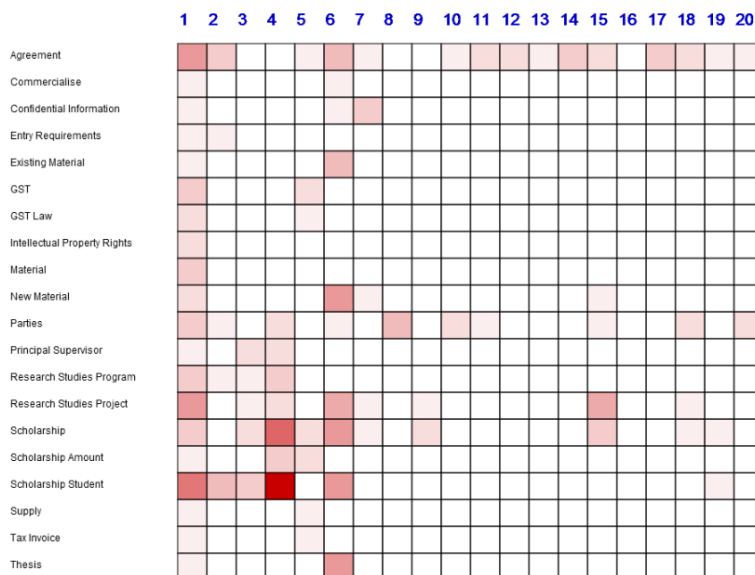


Figure 2.7: Bimodal Representation of a Definition Network in a Contract

As an author of a contract creates it, generates definitions and uses and re-uses them, pathways are created between sections and definitions that are strengthened and extended as the document is elaborated.

Such pathways are also implicit in the navigation of law from section to section. In [Curtotti et al., 2015c][40] a study is undertaken of reading distribution by section in the US Code and US Code of Federal Regulations. That study finds a power law distribution of readership. Sections (nodes) within the legislative network have completely different reading valence - some are read extremely often, others are read extremely rarely.

The conceptualisation of law as strengthening pathways does have a history in the theoretical literature. Holland thought of customary law in precisely such terms.

> *"The best illustration of the formation of such habitual courses of action is the mode in which a path is formed across a common. One man crosses the common, in the direction which is suggested either by the purpose he has in view, or by mere accident. If others follow in the same track, which they are likely to do after it has once been trodden, a path is made. Before a custom is formed there is no juristic reason for its taking one direction rather than another, though doubtless there was*

---

[40]Appendix A.2

*some ground of expediency, of religious scruple, or of accidental suggestion. A habitual course of action once formed gathers strength and sanctity every year. It is a course of action which every one is accustomed to see followed: it is generally believed to be salutary, and any deviation from it is felt to be abnormal, immoral. It has never been enjoined by the organised authority of the State, but it has been unquestioningly obeyed by the individuals of which the State is composed. There can in fact be no doubt that customary rules existed among peoples long before nations or states had come into being. At first no distinction was made between such of these rules as relate to individual character and such as concern society. Morality and customary rules were the same thing, but the distinction between the two was more and more sharply drawn as time went on."* [Holland, 1924, pp57-58]

This account of the emergence of law can readily be related to the modern facts of law. Lawyers engage in behaviours which give rise to "law". Regulators repeat the ritual making of law as a particular way of traversing the pathway from social problem to societal solution. Businesses repeat private behaviours expressed in contracts. Citizens engage with law to manage risks and solve problems. As illustrated by RavelLaw, case law precedents represent repeated behaviours engaged in by people. These behaviours, largely unconsciously, create a network of legal authority. Holland's account is picked up and further analysed by subsequent theorists.

Lon Fuller uses Holland's conception, but emphasises processes of human interaction as a vehicle for norm emergence, rather than unthinking habit.[Webber, 2009]

*"Fuller goes to considerable lengths to emphasize that customary law is not the result of mere habit, mere unreasoning repetition. He takes vigorous issue with Thomas Holland's assertion that customary law develops like a path across a field: people happen to walk on the same line, gradually beating a path into the ground until a right of way forms. Rather, in Fuller's view, customary law is always marked by the need to organize and facilitate interaction. It has a reasoned dimension. Participants perceive the value of norms to their interaction and therefore follow those norms in their conduct. ... Customary law enables participants to coordinate their actions through effective communication, predictably anticipating each other's actions. The perception that an action is obligatory arises – customary law is created – when the participants "have come to guide their conduct toward one another by these expectancies." ... Law is not an abstract theory of justice. It is a method of social organization that is grounded in a particular society, governing relations within that society. It is, in Fuller's words, a "program for living together". ... The members of a society may not share much,*

*but if they are to live in any kind of order they at least need some method, some principles, even if rudimentary, for coordinating their actions."*[Webber, 2009]

In addition to such implications for the processes of law making, network thinking allows us to develop more refined accounts of the boundary of law. We saw above that a primary concern of positivism was to frame a sharp boundary between law and non-law. Critical scholars, on the other hand, sought to remove this boundary. A networked structure to law and related phenomena allow us to ask whether a particular "node" is law, whether it belongs to another set, or whether it belongs to both domains. There is no inherent problem in a legislative statement being both an act of law making and an act within a socio-cultural or political domain - as in the example of the Laws of Aethelberht, which simultaneously "made law" and enhanced status.

Network thinking is also relevant to the application of law to real world networks (noting that the network characteristics of reality are more obvious in some contexts than others). For example, traffic law regulates a traffic network. Similarly, telecommunications law seeks to govern telecommunications networks. Spulber and Yoo consider the implications of graph theory for telecommunications law. They examine the problem of regulating new entrants into a telecommunications monopoly or oligopoly. Regulators have addressed competition by regulating for compulsory access for new entrants to existing telecommunications networks. Generally such new entrants are granted access on the basis of a cost calculated from the opportunity cost of access to a particular element (node) of a network. They show that such a regulatory approach fails to properly account for cost as the true cost of a node depends on its relationship with all other nodes. If a node is on a minimum path through a network, loss of capacity at that node to the original provider may exceed cost, whereas if it is not, there is little opportunity cost to the owner of access granted to a new entrant.[Spulber and Yoo, 2004]

Thinking of law as networks: networks of norms, networks of relationships, networks of behaviours illuminates aspects of the nature of law that are absent from some accounts. Moreover, like thinking of law as language or knowledge, thinking of law as network, facilitates the practical use of law, as illustrated by the examples provided above.

### 2.1.7 Law as Computation

*Jean Valjean: But this is common humanity! Are you a machine?* **Etienne Javert:** *I am an officer of the law doing my duty. I have no choice in the matter. It makes no difference what I think or feel or want. It has nothing to do with me - nothing! Can't you see that?*

Les Miserables, Producer: Cameron Mackintosh, 1983.

Anyone who has been involved in the study of law and the writing of code will draw the obvious parallels between legislation and computer code. Sometimes legislation expressly arises from a desire to "codify" a previously less organised body of law. Computer code is often written as a series of "commands" issued by human "sovereigns" to computational agents. Particularly in imperative programming (as its name suggests) such code is written exactly in this way. Common features of programming such as use of "definitions" ($x = 300$) and "rules" (*if A, then B*), increases the familiarity.

However, more interestingly, for the nature of law, is that computer code need *not* be written as a series of commands. A diversity of programming paradigms (some far removed from notions of "command") have been created by computer scientists.[Van Roy et al., 2009][Thompson, 1999][Chun, 2007][Horstmann, 2008] [Krishnamurthi, 2008] Four prominent paradigms are briefly described below.

1. **Imperative**: Code is primarily conceived of as a series of commands which the human operator issues to the computer.

2. **Object Oriented**: Code is designed as a set of interacting objects. Objects may have characteristics (data) and are able to carry out functions if called by the code. For example a program with a graphical user interface will have programming objects (code which creates virtual objects) which constitute the menus, buttons, windows and other features of the graphical user interface. Such objects may "listen" for events (a click on a button), and respond appropriately.

3. **Functional**: Code is designed as a set of interrelated mathematical functions. Functions are interlinked to produce a mathematical result. The output of one function becomes the input for another in a chain of mathematical operations.

4. **Logic**: Code is designed to answer queries given principles of logic, logical constraints and a knowledge base provided to the program.

Increasingly, such paradigms are used interchangeably in programming languages and as practically needed by developers in a complex and online programming environment.[Krishnamurthi, 2008]

The experience of computer programmers shows a sharp distinction between form and function (in the non-mathematical sense). Programs may be expressed in the form of *commands*, but in reality what occurs in the machine is simply a process of change in an extremely complex electronic register. In legal theory the form, *a*

*command*, is central to how law is defined. The reality of law is a process of change in a complex social register including individuals, communities and institutions.

In Section 2.1.5 I introduced Surden's examination of data-oriented contracting. He observes that "computable contracts" are a possible application of data-oriented contracting.

> *"The basic idea behind a computable contract term is to create a series of actionable, computer-processable instructions that approximate what it is that the parties are intending to do in their contractual arrangement."*[Surden, 2012]

Using this instruction based paradigm, Surden offers examples of contractual terms that could be automated. A program able to compare payments recorded between parties with an obligation to make a payment can compute whether the contractual condition has been satisfied. A program able to parse a date for expiry of an option to purchase can compute whether an attempted purchase under the option has been validly made within the option period. The common theme of such examples is the ability to compare values within a contract with data in the real world. Surden notes that some contract provisions are not so amenable to computation (for example discretionary terms). On the other hand complexity is something that can be captured in computations. He also notes that in a computational context, conveying *meaning* to a computer consists of no more than representing an intended meaning in a corresponding function (operation) in the computer. For example the association of a 'print' command with a printing operation.[Surden, 2012]

Surden's account is highly command-oriented in its account of legal computation. It corresponds to an imperative style of programming.

Another approach to modelling contract terms is to model a contract as a finite state machine.[Daskalopulu, 2000], [Molina-Jimenez et al., 2004], [Flood and Goodenough, 2015], [Anderlini and Felli, 2004]

Flood and Goodenough model contracts as automata - or as finite state machines. A finite state machine has a defined set of *states* in which it can exist. Possible *transitions* between these states are also specified. In the case of contracts, these transitions are associated with *events* in the real world such as actions by the parties. They argue that use of natural language in contracts should be considered in the same light as the operation of a horse in comparison to an automobile. The limitations of natural language can only be overcome by adopting an entirely new form for the expression of contracts. They provide an example of how a well structured financial contract can be a finite state machine (specifically a deterministic finite automata (DFA)). They note that a DFA has no memory of its previous states. All that is required is that the next state can be computed from the current state and from a

set of possible events. For drafters this suggests that this is where attention should be focussed: states, events and possible transitions.[Flood and Goodenough, 2015] Daskalopulu and Molina-Jimenez et al undertake similar work.[Daskalopulu, 2000], [Molina-Jimenez et al., 2004]

As will be evident from the above Surden's explicit conception of the nature of computation and law modelling contracts as finite state machines is quite different, although sharing a vision of automating contract execution. Also it abandons a "law as document" paradigm. That paradigm can for instance be seen in Lauritsen's discussion of software drafting systems for law, in which the role of the software is seen as assisting in the creation of textual documents, texts and metatexts ("the law").[Lauritsen, 2007]

Much work has also been done on computation of contracts and legal instruments from a logic programming paradigm. Such work focusses on developing logical representations that fully capture the meaning of legal instruments. Daskalopulu while using finite state machines also focusses on modelling obligations and interactions between obligations, noting the similarity of her scheme to deontic logic.[Daskalopulu, 2000] This representation is fundamentally concerned with modelling logical constraints. She undertakes work applying logical models to contracts, in extension to application of logic to legislation.[Daskalopulu, 1999], [Daskalopulu, 1998] [Back and Von Wright, 2000] also combine approaches utilising finite state machines and logic programming. [Grosof et al., 1999] use XML and declarative logic for representation of contractual provisions. [Governatori, 2005] represents rules in contracts using defeasible logic. Defeasible logic extends strict logic to enable the representation of soft inferences ("defeasible rules") that are true unless a particular condition ("defeater") causes them to be false. Governatori demonstrates how such soft rules correspond to inferences and defeating rules found in contracts. For example, a premium customer will be entitled to a discount (the defeasible rule), but the discount does not apply if the item is purchased as part of a promotion (the "defeater").[Governatori, 2005] Such situations are commonplace in legal logic.

This brief review of computable contracts illustrates similar intellectual currents in concepts as to the nature of contracts and the nature of code.

What is striking as a conclusion from this review of computation of law, is that *laws are not commands*, or at least if we are to use the concept of command it is better to say that "the law may be expressed in the form of commands". Commands represent a common representation of laws, but command is the form not the essence. Laws, like programs, may be expressed in different forms. When ought commands be used? This is perhaps best considered as a question of design, to a consideration of which we now turn.

### 2.1.8   Law as Designed Artefact

*"**Orestes:** I tell you, the gods should have consulted me before they created any-*
*thing.  **Hypatia:**   Why do you say that, Orestes?  **Orestes:**   It all seems so*
*whimsical.  Why the joint effect of two circles?  Wouldn't it be more perfect if*
*the wanderers didn't wander?  and a single circle gave sense to everything?  ...*
***Hypatia:** I know, I know, but suppose – just suppose! – the purity of the circle*
*has blinded us from seeing anything beyond it! I must begin all over with new*
*eyes. I must rethink everything! ... What if we dared to look at the world just*
*as it is. Let us shed for a moment every preconceived idea – what shape would it*
*show us?"*

Agora, 2009, Director: Alejandro Amenábar

[Curtotti et al., 2015a][41] considers law from the viewpoint of design: indeed as a
designed artefact. This invokes a design paradigm. As observed by Fox: *It is striking,*
*if one has never thought about it, how much of the world around us is designed.* [Fox, 2006]
Law is part of this culture of design. A design paradigm may be further represented
by Norman's observations as follows:

*"[Design] focus[es] on the interplay between technology and people to ensure that*
*the products actually fulfil human needs while being understandable and usable ...*
*not only must the requirements of engineering, manufacturing, and ergonomics*
*be satisfied, but attention must be paid to the entire experience, which means the*
*aesthetics of form and the quality of interaction."* [Norman, 2013]

Design involves meeting a hierarchy of user needs: *functionality, usability and user*
*experience.*[Norman, 2013] and [Haapio, 2013] Designers may adopt a human centred
approach to design and seek to achieve design goals by providing signs (*signifiers*)
and tools (*affordances*) which facilitate use of a tool.[Norman, 2013],[Curtotti et al.,
2015a] For example a door handle implicitly communicates its function to users,
if appropriately designed.[Norman, 2013] The orientation of the process is towards
ensuring that a user's goals can be effectively met while providing a positive user
experience.

In Section 2.1.3 we saw that the primary orientation of drafters of legislation is to
control or influence how legislation is used by a small category of users: judges and
parliamentarians. This corresponds to a focus on designing for functionality. Similar
dynamics occur in contract drafting where legal drafters may be focussed purely on

---

[41]Appendix A.3

legal robustness of a contract and the consequences of drafting failure. The primary audience is an imagined judicial officer who may sit in judgement after a contract has dissolved into a legal dispute. As Haapio points out, this use of contracts (contract failure) is not the typical use case for contracts. The vast majority of contracts are used in their primary context as business and managerial tools.[Haapio, 2013, p27 et seq]

Design perspectives are also influential in the creation of software products. Software design is defined by Fox as follows:

> *"Software design is the activity of specifying the nature and composition of software products that satisfy client needs and desires, subject to constraints."*[Fox, 2006, p6]

Fox suggests that it is helpful to consider this design process as a process of problem solving. The designer is setting out to solve a client's problem.[Fox, 2006, p6] Software product design addresses two levels: product design and engineering design. Product design is "mainly concerned with styling and aesthetics, function and usability, manufacturability and manageability, and social and psychological roles and effects of artefacts and services". Engineering design is concerned with the internal workings of a product - providing those elements that meet the specifications of the product.[Fox, 2006, pp12-14]



Figure 2.8: Australian immigration visa page facilitating use of migration law

An example of how this kind of thinking bears on the communication of law is provided by the online communication of law by government agencies. By and large legal texts themselves are difficult to impossible to comprehend for its average "user". The result is that informational interfaces need to be provided by government agencies to assist users to access the law.

The Australian Migration Act and Migration Regulations offer an example. The Migration Act is 1088 pages long. The Migration Regulations are 2160 pages long. It is a forbidding morass of legislative provisions. A typical user may only need to use the system once, or at most several times during their life. Lawyers and migration agents (usually at significant cost) provide "legal services" to make this system accessible. The Australian Department of Immigration and Border Protection website provides what is in effect a software product in the spirit of Fox's description of software design. Figure 2.8 illustrates the core material presented to a user on navigating to a visa page.[42] Users are assisted to navigate to the page most useful to them by being provided with tabs depending on whether the user wishes to visit, study, live or work in Australia. The 485 visa is a temporary visa allowing graduates of Australian higher educational institutions to remain in Australia to study and work for a period of up to 4 years.

The webpage provides key information on its face: what is the visa for? who might be able to get it? If the visa looks appropriate to the needs of the user, they can then navigate through tabs which provide further details: the application process and the rights and obligations of a visa holder. In addition, the user is provided with tools such as an online portal to make an application through the website. On the left, a user who has successfully applied for a visa, can check their current visa status. There is a lot more going on than the mere publication of a summary of the law. The selection of information and the processes for collection and provision of individualised information is oriented to the concerns of the user. Both the hierarchy of user needs and the framework of product design are illustrated by the example.

A legal "purist" might make two objections. Firstly, that information on web pages is not real law. The "real" law is found in the Act and Regulations. A response to this, is that this is merely a matter of convention. There is nothing that prevents Parliament investing a web page, or its information, with the force of law - although such a step is unlikely to have (yet) occurred in any jurisdiction.

Secondly, a legal purist may object that other functionality provided on the website is not law. A response is that the web pages displace innumerable billable hours that a lawyer or migration agent might otherwise have earnt. From the business of law viewpoint, this is law. Further, from a practical viewpoint, the interactions carried out within the website by the user is the "substance" of the legal process involved. It is what the Parliament wanted to achieve - allowing appropriately qualified individuals to apply for and be issued with appropriate visas. The statute book is relegated to spending most of its time gathering dust on a library shelf.

A slightly different example from the private domain is the creative commons

---

[42]485 Visa web page http://www.border.gov.au/Trav/Visa-1/485- accessed 17 July 2015
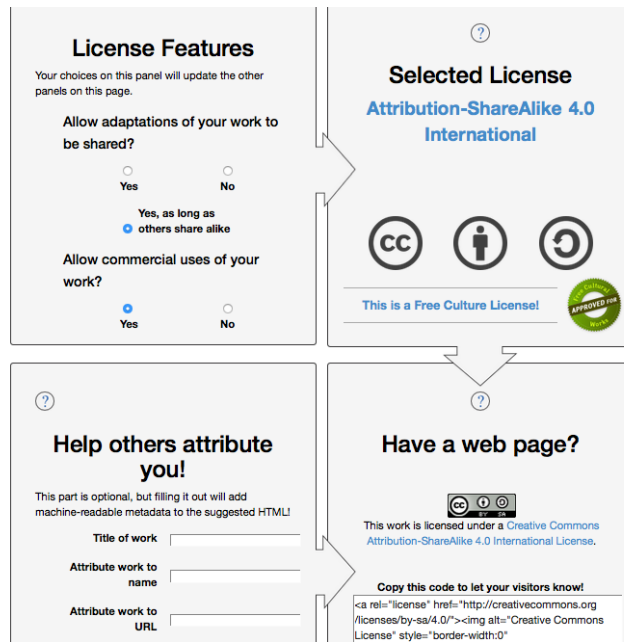
Figure 2.9: Creative Commons Licence creation interface

platform.[Haapio, 2013, p73], [Curtotti et al., 2015a] Figure 2.9 illustrates the process for creation of a complex copyright licence.[43] The process provides a user with simple options and icons to represent the major variables that can be incorporated into the licences. The user can choose whether to allow use for commercial purposes, whether to allow modification, whether to require attribution, whether to require any modified product to also be shared under the same licence. Also they can provide information about the licensed materials which accompany the licence. In addition to providing facilities for licence creation, the system uses visual communication to represent the licence terms. Figure 2.10 shows the icons provided for this representation.



Figure 2.10: Creative Commons Icons

---

[43]Creative Commons Licence creation page http://creativecommons.org/choose/ accessed 17 July 2015.

Both these examples incorporate concepts and approaches found in software engineering. Among these concepts are modularity, abstraction and encapsulation, reusability and polymorphism. Modularity seeks to organise information in well-defined conceptually simple and independent units. Abstraction is the process of ignoring some characteristics so that others can be focussed on - or focussing on essential characteristics most pertinent to the particular interaction. Encapsulation involves hiding unimportant information in a given context. Polymorphism allows a unit within a program to be re-usable within different contexts while retaining a common core functionality.[Fox, 2006, pp 7,233-234, 245, 440] [Horstmann, 2008, 67, 409 et seq] The creative commons licensing platform provides for reuse of the same elements in millions of licences. By 2009 around 300 million licences had been created. ([Kapitze, 2009] cited in [Curtotti et al., 2015a]) Creative commons effectively created an ecosystem of intellectual property licensing that had not existed previously. This combination of software product and legal instrument enabled millions of users to re-use common elements for their particular needs. Non-essential information is hidden from the user.

The immigration web page also illustrates these concepts. The 485 visa page is largely mirrored in similar pages for other visas. Each of them have the same tabs and facilities. Each page is modular (a visa page). Each page hides unimportant information and presents key information as needed by the user. Similar tabs are provided (about this visa, visa application, visa holder). A common core of functionality is being re-used. The information behind the tabs varies according to the particular visa. In some cases additional tabs are provided (for example where an employer is required to sponsor the visa - to provide information and functionality to the sponsor). The linked process for applying for the relevant visa will link to an online form suited to that particular visa. Such characteristics illustrate reusability and polymorphism.

The foregoing are examples of "legal products". Their essential core is the Migration Act and the Migration Regulations. These are entirely hidden from the user - but key information is abstracted and presented. The law is however complemented with other features which deliver the overall product (a process to find and make use of the law relating to issuing of a visa).

Similarly to our consideration of law as computation in Section 2.1.7, the character of law as "commands" largely drops away from a design perspective. Neither the immigration or creative commons examples are concerned with control. Rather they implicitly embody a paradigm of empowerment.

The implications drawn to our attention by this "designer's view" of law relate to the theoretical discussions of the nature of law and power which have been explored

in Section 2.1.2.5 and the broader theoretical discourse concerning the nature of law. Laws as commands suggest an apex at which sits a sovereign. Where law is critiqued as the use of power to impose the will of the powerful on the weak, this implicit model still sits in the background. A design perspective is informed by entirely different assumptions. The user is at the apex and the concern of the designer is to meet the user's needs. It is an inversion of the command theory of law.

The feminist discourse around the nature of power also bears on a design perspective. Feminist attention to "power to", "power with" and "power within" all relate to the creative aspects of design. We can also apply the Karlberg bi-dimensional mapping of power to the design space. Of course not all designed artefacts are primarily concerned for the welfare and aspirations of human beings. A cigarette is a designed artefact, yet the designers ignore important health needs of their users. In fact, the design in that case is predatory. It seeks to entrap the 'client' into behaviours that meet not the needs of the client but the needs of the designer. Weapons may protect some human beings, but cause considerable harm to others. This spectrum of potential design "motives" corresponds to how power may be used: adversarially at one end to cooperative at the other. Similarly a designer may have different power relationship to a client. A client may be more, equal or less powerful than a designer. Design can thus fall anywhere in the Karlberg mapping. However, we see that its instincts are on the right hand side of Figure 2.1, and frequently in the lower right quadrant.

The three level hierarchy of design values of function, usability and user experience also draw us to a consideration of these elements in the law.

Summers looks at the functions of law in *Law: Its Nature, Functions, and Limits*.[Summers and Howard, 1972] He does so as an educator concerned to teach law as part of a general education in the humanities. He examines the nature of law from the viewpoint of how it functions. In his analysis, the nature of law may be discerned from *how* law functions. Law may function as a grievance remedial instrument (e.g. as in the law of tort); it may function as a penal instrument; it may serve as an administrative regulatory instrument; as an instrument for organising the conferral of government benefits; or as an instrument for facilitating private arrangements.[Summers and Howard, 1972]

Summers' broad treatment of the functions of law may be contrasted with the comparatively narrow consideration of function undertaken by Hart. This narrowness may be discerned from Hart's focus on investigating the nature of rules and the following statement in his seminal work *The Concept of Law*:

> *"... as we have seen, there is a perennial danger of exaggerating these [affinities*

*with morality] and of obscuring the special features that distinguish law from other means of social controls."* [Hart, 1961, 208]

In other words, from Hart's perspective, law is about control. This is only one function of law in Summers' schema, which provides a more accurate and complete description of law. Summers continues his analysis in respect of *what* functions the law may have. In this respect, he investigates functions of law in promoting human and environmental health, family life and privacy, a peaceful community, basic freedoms, equality of opportunity, protection of private property and supervision of the use of power.[Summers and Howard, 1972]

Control moreover, can be, but is not naturally inherent in a design approach to law. From a design viewpoint, law may be as much concerned with empowerment as with control. Where, further, command or control is used in law, it is from a design viewpoint functional. The design goal is not control - it may be any of a range of societal outcomes to which the law is applied.

## 2.1.9  Law as Emergence and as Complex Adaptive System

*"Oh, God of dust and rainbows, help us to see, That without dust the rainbow would not be."*

From *Two Somewhat Different Epigrams* The Collected Poems of Langston
Hughes, 1994.

As we have continued our investigation of the nature of law through these multidisciplinary lenses and the "law as ..." framework, it becomes increasingly evident that simple crystallisations or reductive definitions of law are unhelpful. Particularly so, if they are taken to represent a complete account of "the law". Such accounts oversimplify. The law has a degree of complexity that defies reductive definition. The complexity of the nature of law suggests that a consideration of complexity theory may offer further insights into the nature of law. The following explores complexity in application to law, including through a consideration of the implications of the phenomenon of emergence.

In 1970, when computer resources were still scarce, John Conway developed a mathematical game involving simple cellular automata. Essentially, he imagined an infinite grid in which a square could either be alive (black) or dead (white) and in which the state of cells would be updated in discrete time periods. Simple rules govern cell transitions in this two dimensional world.

**1.** If a cell is alive and has 2 or 3 neighbouring live cells it continues to live in the next time step.

**2.** A live cell with four or more live neighbours or 1 or fewer neighbours dies.

**3.** An empty cell with 3 neighbours comes to life.[Gardner, 1970]

This is of course a simple set of rules ("commands"). Counter-intuitively it is capable of generating enormous complexity and a diverse menagerie of 'creatures'. Its diverse manifestations (some of which are self-reproducing) have been given nicknames such as 'gliders', 'glider guns', 'beehives', 'clocks', 'pulsars' and 'puffer train'.[Gardner, 1970], [Bays, 1987], [Schulman and Seiden, 1978] Figure 2.11,[44] provides a snapshot of a time step in such a simulation. From simple rules, a whole world of complex entities and interactions emerge. This artificial world sparked serious scholarly study including as a system of artificial life,[Langton, 1986] for the study of physical stochastic systems,[Schulman and Seiden, 1978] and for extension to three-dimensions.[Bays, 1987]

In Section 2.1.3, we introduced formal languages used by computer scientists for the precise expression of software. We are all now familiar with realistic computer graphics that create convincing virtual worlds. A particular variety of formal language has contributed to this realism. Lindenmayer systems, initially developed to model the growth of multicellular organisms were later applied to modelling the growth of plants. Unlike the formal grammars described above, Lindenmayer systems have only terminal symbols. Essentially, a language of three symbols can be used with the Lindenmayer grammar to produce life like plants. Again we see simple rules producing what we experience as complex results.[Smith, 1984] In this case there is a link between complexity and concepts of communication (formal languages) we explored earlier.

A further example of simple rules giving rise to life-like complexity is Reynold's Boids. In this case, the rules simulate flocking behaviours in birds and fish. Here many agents without central authority are able to coordinate their actions to bring about coherent flocking. The rules, at agent level, are as follows:

**1.** Avoid collisions with near neighbours.

**2.** Seek to match velocity with neighbours.

**3.** Seek to stay close to neighbours.[Sipper, 1995]

---

[44]Image Source: Wikipedia creative commons
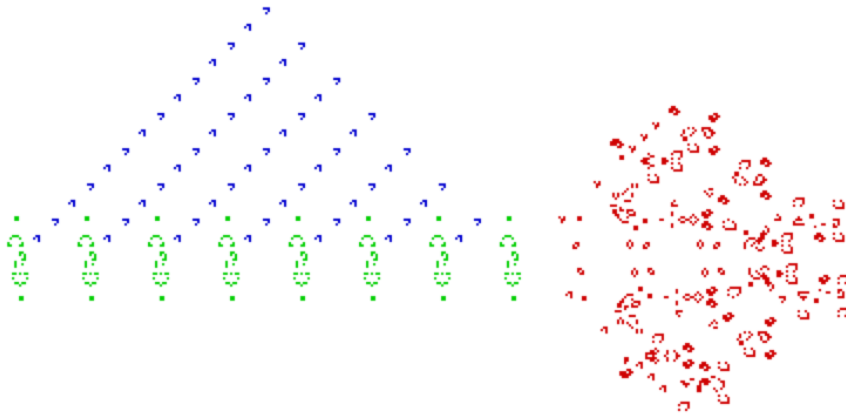
Figure 2.11: John Conway's Game of Life

We have 'commands' but no 'sovereign'.

Reynolds boids are a subset of swarm behaviours that potentially give rise to "intelligence" - as an emergent characteristic of simple rule based behaviours of agents. Insect behaviour in insect swarms has been studied as a way of solving optimisation problems. Such problems can be extremely difficult, requiring finding an optimal solution in a search space which is enormous and which requires a prohibitive computational time to search. Blum and Li discuss how simple rules for ant behaviour in an ant colony behaviour can be used to provide an approximate solution in computationally realistic time and catalogue the wide range of applications of ant colony optimisation. For example, scheduling problems, vehicle routing problems, graph colouring, data mining, bioinformatic problems, and others.[Blum and Li, 2008]

These examples, although embodying 'rules' are distant from the social context of the law. Emergence of norms among social agents as "habit", was studied by Alexrod in the 1980s. Alexrod undertook an investigation of norm formation through an evolutionary computer simulation. He was concerned to study how norms emerge among social agents where a central authority is lacking.[Axelrod, 1986] He defined norms as follows:

> *"A norm exists in a given social setting to the extent that individuals usually act*
> *in a certain way and are often punished when seen not to be acting in this way."*
> [Axelrod, 1986]

We see echoes of both Hart and Kelsen in this definition, although for Alexrod the existence of a norm is not a black and white proposition - a norm has a degree of existence in Axelrod's model. Axelrod creates an environment in which digital

agents play a variety of the prisoner's dilemma.[Axelrod, 1986] In its classical formulation, the prisoner's dilemma involves a situation where two agents play a rational game and seek to choose the optimise their situation. Both gain somewhat if neither cheat, one gains more and the other loses more if both cheat, and both lose more if both cheat. Neither agent knows whether the other will cheat.[Axelrod and Hamilton, 1981] In Axelrod's formulation, the agents may gain rewards by cheating and suffer loss by being punished if caught cheating. In this case the agents have a probability of being observed if cheating. Also, the agents' behaviours are modified through an evolutionary algorithm, as the "game" is played multiple times. This can be interpreted (in one version) as a process for learning behaviours most likely to be beneficial to the agent, in a context where the agents observe each others previous behaviour. Agents may be bold or vengeful and their behaviour evolves as successful agents are rewarded by allowing them to reproduce more effectively in subsequent generations. Finding that norm maintenance was fragile, Axelrod experimented with including metanorms in his model. In this version those who failed to punish a cheater would also be punished. Such metanorms helped to maintain a stable norm environment.[Axelrod, 1986]

Another more well known aspect of Axelrod's work moves us away from a solely punitive conception of norm emergence. In *The Evolution of Cooperation*, Axelrod studies the emergence of cooperation in the context of the prisoner's dilemma. This dilemma is essentially based around options facing a decision maker who can obtain additional benefits from cooperation, but risks harm if the other party does not reciprocate. The basic set up of the game is that the sum of benefits available through cooperation is greater than the maximum benefit available to a single player in the context of non-cooperation. A simple "tit-for-tat" rule emerges as the most successful strategy in this investigation. Cooperate first and continue to cooperate, but if the other party cheats, cheat against them next time. Axelrod suggests the insights of this 'game' carry implications for current international order which is characterised by "egoistic" agents (states) interacting in an environment where central authority is lacking.[Axelrod and Hamilton, 1981] In the context of his work, punishment only makes sense as a strategy supporting a generally cooperative environment. A legal system is as importantly influenced by agreement or reward, as it is by punishment. To conceive of laws only as 'punitive' is to miss this alternative modality in which the law can and does function.

In the examples above, we have seen simple rules given rise to unexpected and complex results. Such characteristics which are not inherent in the rules which give rise to them are often referred to as "emergence". The relationship between the rules and the resulting behaviours is far from intuitive. The law, like other phenomena

displays emergent features. Emergence is described by Goldstein in his review article *Emergence as a Construct: History and Issues* as follows:

> *"Emergence ... refers to the arising of novel and coherent structures, patterns, and properties during the process of self-organization in complex systems. Emergent phenomena are conceptualized as occurring on the macro level, in contrast to the micro-level components and processes out of which they arise."*[Goldstein, 1999]

Emergent phenomena have characteristics such as radical novelty (emergent features are not predictable in component parts); coherence (they manifest within a larger whole); emergence is concerned with macro rather than micro characteristics of phenomena and emergence is a dynamic process. Emergent phenomena cannot be reduced to the sum of the parts from which they arise. They are related (but not identical) to thinking which prioritises 'the whole' over its parts.[Goldstein, 1999] Goldstein notes that emergence plays a particular role in providing explanation or description of phenomena where micro-characteristics cannot. In other words 'emergence' is valuable in providing a vehicle for describing and analysing what is outside a reductive analytical frame. Implicit in the approach is that a description of reality need not be confined to one level. It implicitly rejects what Goldstein calls "ontological-level monism": that reality can only be described at its most basic level. It thus carries implications for the philosophy of science, in that it welcomes study of phenomena at a plurality of levels.[Goldstein, 1999]

An example of emergent phenomena in the law, can be seen in the continuing and unpredictable emergence of new bodies of law on the 'tree' of law from initially very simple beginnings. Intellectual property law, for example, has this characteristic. From initial relatively straightforward rules for the protection of inventions or authorship emerging some centuries ago, a complex legal sub-system has emerged which applies increasingly novel forms of interacting property law, to human creativity and ingenuity. For example, software may have associated with it patents, copyrights over different components, trademarks and confidentiality. We can also see the emergence of today's complex traffic regulations from the interaction of human agents over time. What might initially have started as a path in a meadow may now be a superhighway. As it has developed, a body of law has emerged to regulate it. The emergent characteristics of the system as a whole result in a functioning transportation system. These characteristics are not found in the individual rules or physical structures that make it up.[45] It is also interesting to consider the future of traffic "rules" as driverless cars enter, and in all likelihood ultimately dominate, the traffic system. Might the rules of the road in a fully automated environment come to

---

[45]See also discussion above of Fuller's work on pathways and custom in the law.

resemble something more akin to the rules of Reynold's boids, rather than the complex regulations with which we are today familiar? These rules after all are simply heuristics for optimising traffic flow while maintaining safety.[46]

In section 2.1.6 we explored the network characteristics of law. We saw the pervasiveness of networks in law. We also saw in Katz and Bommarito's work, that in one case the complexity of a legal network increased over time. Further, we have seen the complex (in the sense of many hued) character of law, both in the theoretical background that has been discussed and in the practical investigations connected with this thesis. These factors taken together suggest an exploration of a complex systems view of law.

An approach grounded in complex systems is almost entirely absent from core legal jurisprudence as described in overviews of that jurisprudence.[Doherty, 2002], [Cotterrell, 2003], [Leiboff and Thomas, 2004], [Veitch et al., 2007][47] There is however a specific body of scholarly thinking around law as complex system - particularly as a complex adaptive system. We briefly review this below to illuminate the nature of law as such.

Over time, the complexity and number of regulations and law increases.[Bommarito and Katz, 2010], [Ahdieh, 2006] [Ruhl and Salzman, 2002] Ahdieh notes the emergence of regulatory dependence and regulatory overlap where multiple regulatory agencies find themselves caught in patterns of interactive complexity. Ahdieh refers to this as intersystemic regulation and notes that understandings of law which focus on 'law as rules' have tended to minimise attention to such effects.[Ahdieh, 2006] Essentially Ahdieh here describes the reductive patterns of analysis to which Goldstein refers also. Ruhl and Salzman also discuss the problem of regulatory accretion referring to a range of metrics establishing the empirical reality of the issue.[Ruhl and Salzman, 2002]

Jones begins a review of the scholarship on law as complex system by noting the pervasiveness of complexity and emergence in a wide range of domains from the flocking of birds, to the progress of forest fires. In respect of implications for public

---

[46]A report on the future of automated transport issued by Western Australia notes that:

> *Full automation will ... result in infrastructure changes such as the removal of traffic control. Static and digital signs may become redundant, although there will still be challenges in providing a controlled environment for vulnerable road users to cross the road.* Automated Vehicles: Are we ready? Internal report on potential implications for Main Roads WA https://www.mainroads.wa.gov.au/Documents/Automated%20Vehicle%20Report.RCN-D15%5E2381741.PDF accessed 6 October 2016

Goodall considers the need to program automated vehicles for ethical decision making. Strictly following road rules does not guarantee harm minimisation.[Goodall, 2014]

[47]A brief reference to law as system appears on page 8 of Veitch, Christodooulidis and Farmer's work, but does not refer to complex systems theory.

policy in addressing racial discrimination and resulting segregation, he notes the non-linear effects of different policy prescriptions. He observes that legal scholars have begun to study such effects. He notes the already widespread scholarship and notes its many contributors, particularly identifying Ruhl a leader in the field.[Jones, 2008]

Ruhl notes that despite the endless literature concerned with the "legal system", few legal scholars "have anything to say about [the law] as *system*".[48] Areas of neglect include defining 'what' it is; how it behaves, what are its boundaries; and how does it evolve over time. He notes further that many authors note the complexity of the law, but take the observation no further. He applies complex adaptive systems theory to attempt such a description. Ruhl situates law as a complex adaptive system interacting in complex ways with itself and other complex adaptive systems in which it is embedded such as economics and society. He defines complex adaptive systems theory ("C.A.S. theory") as the study of interactions among agents and the "aggregate product of their interactions". Other agent based analyses either deal with a situation of an unrealistically small number or an infinitude of agents. They are unable to study interactions of moderate scales, something which complex adaptive theory addresses. Complexity is characterised by a multiplicity of interactions among elements, for example such that the removal or change in one element may cause widespread systemic effects. Complex adaptive systems are comprised of heterogenous interacting agents. Deterministic rules determine behaviour at micro-level, but give rise to non-linear effects at system level. A characteristic of complex adaptive systems is their *path dependence*: the past limits the possibilities for future evolution of the system. Other characteristics include "resilience", "stable disequilibrium" (i.e. a stable but constant state of evolution), but also that such systems experience "phase transitions" - unpredictable points at which a system will move from one state of "stable disequilibrium" to another. A prominent concern in Ruhl's analysis, is the implications of C.A.S. theory for the design of law.[Ruhl, 2008] Nicolis and Nicolis describe the signatures of complexity as emergence and the "intertwining ... of large scale regularities and ...of seemingly erratic evolutionary events". Complexity is characterised by multiplicity and a capacity to adapt, evolve and a "capacity to choose between" possible outcomes. Further relationships in complex systems are non-linear. Finally such complex systems may transition between states.[Nicolis and Nicolis, 2007, p4] Boccara suggests three characteristics of complex systems: emergence; consisting of "a large number of interacting agents" and the absence of a central controller determining emergent featuers.[Boccara, 2004, p3]

---

[48]This observation is consistent with insights from Goldstein concerning ontological-level monism: systems level analysis is excluded from the generally accepted frame of legal scholarship.

In respect of the law itself, Ruhl observes that all the typical characteristics of a complex adaptive system are found in the legal system. Ruhl does not assert 'proof' that law is a complex adaptive system, but does assert that the features of law make C.A.S. theory a framework for thinking about the design of law. He emphasises three points in this regard. It is not possible to isolate a change in a legal system. Unintended consequences of change are unavoidable. Accordingly design should be based on C.A.S theory. This also suggests, he says, potential research directions, particularly in respect of the dynamics of legal systems - how they change over time.[Ruhl, 2008]

Ruhl's treatment is not necessarily intuitive - but it can best be seen in his references to the evolution of environmental law. Similar observations can be made about any area of law. Above, the example of intellectual property law is cited. International law has shown a similar growth and transition from relative simplicity to evolved complexity over the course of four centuries when the basic framework of the international legal system were set by the Treaty of Westphalia. Initially a limited set of principles governing interacting nation-states as the sole actors of the system, it has evolved into a complex system of interacting legal domains and regional and international institutions, which now profoundly interpenetrates the domestic legal domain.

Ruhl and Salzman also draw attention to design issues in the context of the growth of the modern 'regulatory state'. Rules may be inefficient (their costs to society may exceed their benefit or they may divert resources from socially valuable applications). Rules may be complex: "agencies often produce rules that are complicated, difficult to understand, ambiguous or contradictory". Agencies involved in rule making lack sufficient democratic accountability. Finally, the modern regulatory state has seen a process of regulatory accretion (increase in the number of regulations over time). They suggest that the traditional approach of trying to fix 'individual rules' is not a useful response to the problem of regulatory accretion. The problem is systemic and needs to be addressed at system level.[Ruhl and Salzman, 2002] They note that "it is not so easy to identify which rules to revise or remove so as to eradicate system burdens that contribute to non-compliance or other system failures".[Ruhl and Salzman, 2002]

In [Curtotti et al., 2015c], however, it is shown that only a small proportion of regulatory language is likely to represent most of the regulatory burden, as a small proportion attracts the bulk of readership.[49] Nonetheless, they have a point. An interconnected complex system presents costs beyond the cost of compliance with any single rule. Complexity grows exponentially with growth in interconnected compo-

---

[49]See discussion in Section 4.1 of Appendix A.2.

nents. The difficulty with designing at system level however, is that the emergent properties of the system can only be appreciated by observing the system in action. One manifestation of complexity is involuntary non-compliance through regulatory overload or regulatory ignorance as regulation exceeds the legal resources available to most economic agents. At a systemic level a response employed by many regulatory agencies is *assisting* compliance by *"compliance outreach, hotlines, plain English translation of rules and the like"*. Ruhl and Salzman encourage regulators and regulated to think systemically. For regulators, to make greater use of feedback mechanisms. For regulated to use management systems to address compliance risks.[Ruhl and Salzman, 2002]

Hathaway focusses on the path dependence (a characteristic of complex adaptive systems) stating: *"it is impossible to understand the law as it is today without understanding the law as it has been in the past"*.[Hathaway, 2001] Path dependence implies that the future characteristics of law are determined by its past. Path dependence implies the difficulty of maintaining an optimal legal system and the unpredictability of future outcomes. Further, it implies that opportunities for effective change occur only rarely in the system. *"The order in which cases arrive in the courts can significantly affect the specific legal doctrine that ultimately results."*, she states. Such observations are unsettling to those believing in the rationality of law. She distinguishes three forms of path dependence: increasing returns path dependence; evolutionary path dependence; and sequencing path dependence. "Increasing returns" dependence arises where as a new pattern becomes established, there are gains in continuing to reinforce that pattern - so that it comes to dominate. (For example the dominance of the QWERTY keyboard can be explained in this way even though it is known to be sub-optimal.) In respect of "evolutionary path" dependence, Hathaway notes recent theory in evolutionary theory that support a pattern of evolutionary development that is marked by long periods of little change with brief periods of rapid change. The periods of rapid change involve sudden diversification of species. "Sequencing" path dependence arises from the impossibility of multiple choices and multiple choosers being in the same position on a choice pathway. The choices made by one actor determine the choices available to subsequent actors. Choice sequence is arbitrary - and because of its implications, unfair. She applies this theoretical framework to analyse path dependence in the law - particularly in the context of the common law doctrine of precedent. Understanding path dependence assists in analysis of processes of change and stability in law. Further it has implications for the doctrine of precedent itself.[Hathaway, 2001]

Holz investigates the implications of complexity theory for judicial decision making in the context of Heisenberg's uncertainty principle and Godel's theorem. The

former establishes a limit to human capacity to measure reality. At a microscopic level our picture of reality becomes uncertain. Godel's incompleteness theorem established that no formal system is able to establish both its completeness and the truth of all its propositions. In describing complexity theory, Holz also notes the concept of *attractors* (end points towards which systems evolve). Such end points can be fixed, cyclical or "strange" (the last known as "strange attractors"). Unlike the other two types, strange attractors do not repeat nor reach a stable point. The initial conditions greatly influence outcomes, which are largely unpredictable in such systems. Holz applies these concepts to consider judicial decision making, concluding that the view that judges carry out a deterministic application of positive rules does not correspond to reality; and similarly that nominalism (judges ought seek a just outcome in the case before them) represents an extreme. He favours judicial realism as best adapted to addressing the complexities of legal decision making, which by its nature seeks to determine a simple end point from extremely complex inputs.[Holz, 2007]

LoPucki is also concerned to apply a systems approach to law. A systems approach concerns itself with the whole, as much as the parts. It resists the analytical approach of simplifying a phenomenon in order to analyse it. She adopts the following definition of system: *"a regularly interacting or interdependent group of items forming a unified whole"*. Units within a system may be human beings, atoms, or systems themselves. In contrast to the research cited above, she adopts "systems analysis" from the domains of engineering, business information systems and computer science. Systems analysis seeks to analyse a system in terms of its components parts and their interrelationships. System analysis involves a number of discrete steps: identifying the system; identifying or attributing goals to the system; identifying the component parts and functions of the system; describing the relationships between the component parts; identifying inconsistencies between goals and functioning of the system. Attribution of goals may be empirical (the system has these goals); or normative (the system should have these goals). [LoPucki, 1996]

Like Hart, LoPucki considers law to be primarily a means of "control". In this context it is related to other "control mechanisms" such as social or physical controls (for example barriers). Taking this approach, she defines the "legal system" as law using mechanisms of social control (as opposed to methods of control not using laws). She provides a number of examples of application of systems analysis to problems of law. She cites her own work in improving copyright filing systems and applying systems analysis to comparative study of US and Canadian legal systems. She notes examples where substituting physical systems of control (for example electronic payment vs. legal liability for funds behind cheques) means that it is useful

to sometimes look outside "law related" mechanisms to achieve the same goals. A systems analysis approach is heavily empirical.[LoPucki, 1996]

The work cited above is often concerned with question of design in a complex system context. Johnson examines this nexus (one which extends to all complex human artefacts). Historically, science addressed problems which were either intrinsically simple or which could be be simplified so as to be usefully investigated by reductive analysis. Problems were defined in advance and could not be changed. Science now increasingly addresses complex problems which cannot be solved in that way. Complex science faces the problem that it seeks to make predictions about complex systems. Yet such systems are computationally irreducible. Only by running the system can its future state be known. (Conway's Game of Life, being an example). Complex systems are often associated with human imposed goals - desired future states for systems. Johnson discusses the relevance of design to these characteristics. Design he observes, can be defined as a process in which problem and solution iteratively emerge from each other in the design process. He suggests complex scientists should seek to work with policy makers in studying large scale interventions in complex systems. Both computer simulations and study of real phenomena can offer insights into the behaviour of complex systems.[Johnson, 2010]

The arguments that law is a complex adaptive system are persuasive. Yet this brief exploration suggests that the body of knowledge concerning the implications of this observation are still nascent. How the law might be usefully approached with this insight in mind is still uncertain, despite the growing body of research on the law informed by systems thinkings. It certainly assists us in appreciating that simple analyses of law are inadequate. Yet we are left with the reality that an implication of complexity theory is inherent uncertainty and unpredictability in any intervention in the system.

Nonetheless the observation helps us understand some of the puzzles concerning attempts to enhance the communication of law (the central concern of this research). Centuries of effort have been devoted to the question - yet progress seems a case of one step forwards, two steps back. A case in point is the (at the time) revolutionary reforms in legal writing of the Victorian era. These reforms introduced 'innovations' such as paragraphing, definitions, section numbering and professional drafting offices.[50] Yet some of these very reforms are likely now implicated in the current difficulty of legal language. Definitions give rise to complexity through complex interlinked definitions that potentially expand the language of a rule by orders of magnitude.[51] Sentence length in legal sentences blows out to hundreds of words -

---

[50]See discussion in Section 2.2 of [Curtotti and McCreath, 2012] (Appendix A.6).
[51]See discussion in Section 1 of [Curtotti et al., 2013] (Appendix A.5).

disguised and facilitating by paragraphing which gives a superficial appearance of simplicity.[52] Solutions to old problems give rise to new and unanticipated problems. The pattern is characteristic of complex adaptive systems.

Language itself, it might be noted, is a complex adaptive system.[Beckner et al., 2009] Accordingly, law *as language*, is also a complex adaptive system in respect of that language.

Like other C.A.S's language is characterised by interaction among a multiplicity of agents (speakers); it adapts and evolves based on past speaker behaviours; and language emerges from *"interrelated patterns of experience, social interaction, and cognitive mechanisms"*. Language is an emergent property of local interactions between speakers whose actions are independent and local rather than centrally coordinated and global. There is no ideal language speaker - each speaker contributes to the emergent characteristics of language. Languages are perpetually dynamic. Competing factors such as speaker desire for economy and hearer desires for clarity of communication contribute to the emergent characteristics of language. Like other C.A.S. languages are subject to non-linearity and phase transitions. Language is sensitive to the social networks in which it emerges.[Beckner et al., 2009]

Historically the primary agents of legal language have been the legally trained. This long established pattern is being disrupted as a result of information technology. Non-lawyers are now a substantial if not majority audience which read law online.[Curtotti et al., 2015c][53] Further, we may note that the production of legal language will not in future be confined to human agents. Increasingly, legal language will be produced by computational agents. (See discussion above in Section 2.1.7) Such changes suggest that we may see a phase transition in the nature of legal language. The demands and impacts of these new users for effective communication in the complex adapative system which is legal language will change its nature, although the nature of the future form of that language may be currently unpredictable.

An important driver for how law is written, is the assumption that it is possible for the writer to control, by careful drafting, how the law will be used. Confidence in this belief is significantly eroded by complex systems theory. Appeals to the necessity of precision in drafting look increasingly hollow and counterproductive.

Looking at law as complex adaptive system has drawn our attention to the dynamic processes of law, and continues to affirm the inadequacy of conceptual models based on reductive definitions of the law. Yet even complexity theory itself is inadequate - as we see that it is incomplete without inclusion of other perspectives of what

---

[52]See discussion in Section 5 of [Curtotti et al., 2015b] (Appendix A.1).
[53]See discussion in Section 4.3 of Appendix A.2.

the law is. For example, it makes little sense to ignore that the raw subject matter of the law is well expressed by the 'law as rule ...' paradigm.

### 2.1.10   Conclusion

The foregoing discussion has considered law from a multidisciplinary point of view adopting a "law as ..." paradigm to seek to develop a better answer to the question: "What is law?" The approach suggests that the law can be liberated from false dichotomies that demand that law be understood only through the lens of a particular theory. Further, tracing this multidisciplinary journey illuminates the narrowness of scholarship that sees law almost exclusively through the lens of "rule".

In reviewing some of the dominant theories of the nature of law we saw that they tended to share a common perspective that the law is *essentially* a set of rules. Dominant among these theories is positivism. Yet even those which challenge positivism generally begin from the same rule focussed premise. This approach has much to do with the more general rise of positivism in scientific thought. This thought, asserted itself as rational, objective and values-free. Science has been extraordinarily successful in the class of problems to which its traditional analytic tools are most applicable.[Mingers, 1980]

As Nicolis and Nicolis describe it (referring to the Newtonian paradigm):

> *"... science is usually viewed as an algorithm for predicting, with a theoretically unlimited precision, the future course of natural objects on the basis of their present state ... the world is reducible to a few fundamental elements animated by regular, reproducible and hence predictable behaviour: a world that could in this sense be qualified as fundamentally simple."*[Nicolis and Nicolis, 2007, pp1-2]

This pattern is fully replicated in positivist theories of law. Law is found empirically in the actions of 'the sovereign'. It is moreover certain in its character and deterministic in its application.

Realists, as we saw, proposed skepticism about the certainty of both the rules and the facts to which they are applied. Feminists problematised the objective-subjective dichotomy and gender bias. Critical scholars queried the relationship between power and law and feminists and others queried the nature of power. Such theories while still grounded in "law as rule" begin to challenge the certainties of positivist law. New theories of power (with which law is intimately connected) open the potential for new understandings of law. From traditional theories we learn particularly about law as rule. We understand the distinction between law, so defined, from non-law

with which it is connected. That is, related systems of social control such as morality or social norms. Or alternatively we argue that the chasm is illusory. We see also the strong connection of legal theory with that which was the 'state of the art' in scientific thought of the late nineteenth century.

As multidisciplinary perspectives were applied to the nature of law - each metaphor of law illuminated new aspects of legal phenomena. In turn, we investigated law as communication; law as document; law as data, information and knowledge; law as network; law as computation; law as designed artefact; and law as emergence and complex adaptive system.

As document, law has been both emphasised and ignored, depending on the context. We saw also that this 'form' in which law is captured unnecessarily constrains the utility of law. The law can provide new outcomes if conceived in new forms, for example, as data. Examining law as computation allowed comparisons to be drawn with the field of software engineering. This field's use of 'commands' but equally adoption of other programming paradigms pointed to the conclusion that 'command' is not essential to law - it is a form in which the law may (and of course often) is expressed. The conception of law as a form of control has been dominant in legal scholarship. We saw that this conception was challenged by approaching the law as design. The principal function of design (often) is not to control, but to empower. It is possible to design laws to empower rather than to control. This observation evokes the feminist discussion of the nature of power and how it is used. We saw that at least some scholars have a broader conception of law - exploring its functions other than the function of control. A future research direction would be the investigation of law as empowerment. Considering law as network highlighted law as pathway and connection.

We have seen that relationships between metaphors of law are not isolated. Law is communication and language. Language is a complex adaptive system. Law, also, is a complex adaptive system. Complex systems, as in Lindenmayer systems may be expressions of a formal language. The law may be 'computed', as may complex adaptive systems. The law is designed. Language also, is to some extent, designed. We may think, for example, of the human construction of 'national' languages which suppress local variance and emphasise difference with other 'national' tongues. These various metaphors of reality are not disjointed, separate spheres. Rather they are interconnected with each other in a complex network of relationships.

We may note further that although language and law have much in common (in some respects law being a subset of language), our mental models of their dynamics are extremely different. Language (largely) arises organically from the communica-

tive acts of individuals who each contribute to the emergent properties of language as a complex adaptive system. The command theory tends to strip away such realities from 'legal communication' - yet like language - the law arises from innumerable interactions of human beings about the law - an individual deciding to sue - two parties entering a contract - a community organisation agitating for change to the law - legislators interacting in the process of making law - agencies considering how better to achieve their purpose.

The scientific inspiration of modern theories concerning the nature of law, particularly legal positivism, has also been described above. Science, however, has moved on. It is appropriate that our models of law, while avoiding a crude 'scientism' of the past, take heed of the revolutions in scientific paradigms that have occurred since the early twentieth century. Indeed the words 'paradigm' and 'revolution' are terms that are themselves part of that history of transformation. Legal positivism is part of a broader positivist movement beyond the legal sphere that sought to distinguish metaphysics from empirical investigation and which sought to promote "scientific inquiry" as *the* way all intellectual enquiry must be undertaken. An example is the Logical Positivists in 1920's Vienna.[Ray, 2000]

At this point it is illuminating to briefly examine increasingly nuanced understandings of the nature of science. The following paragraphs draw on Chalmers' book: *What is this thing called Science?* Karl Popper and his successors who took the Logical Positivists as the point of departure, began a reconceptualisation of science that has moved a considerable distance from positivist constructs of science. Interestingly, given the character of legal positivism, which sought to demarcate law from non-law, Karl Popper was deeply concerned to develop reliable criteria for 'demarcating' science from non-science. Popper highlighted the inadequacies of the idea that scientific knowledge can be derived solely from induction based on observable facts. Given this, how was science to be distinguished from pseudo-science? Whether a theory could be falsified was the criterion he adopted for demarcating science from non-science. A theory cannot be scientific unless it is in principle possible to falsify it. Science progresses when its previous theories are falsified by new experiments.[Chalmers, 2013, pp 3, 55, 67, 68, 94] The problem with various versions of this falsificationist view of science is that, if scientists really behaved that way, science would rarely progress. An examination of historical examples, such as the emergence of the Copernican revolution, show that it took centuries of investigation, during which time the "observational evidence" and existing theory supported the conclusion that Copericanism was false. Scientists however, did not abandon it.[Chalmers, 2013, p 81 et seq]

Regarding Popper's formulation as inadequate, Thomas Kuhn, who has been

influential in the post-war period, introduced notions of scientific paradigm and scientific revolution. For him science, typically lives within a "paradigm" - a scientific world-view - which defines the boundaries of legitimate scientific endeavour. In contrast to Popper, scientists are not engaged typically in 'falsifying' their science. Rather they are engaged in puzzle-solving - elaborating, deepening and strengthening the existing paradigm (what Kuhn called "normal science"). Scientists moreover form scientific communities who support and sustain the paradigm. From time to time normal science brings to light irresolvable problems that are inadequately explained by the paradigm. Because of the success of the scientific paradigm, these problems (falsifications from a Popperian viewpoint) may be lived with for some time rather then bringing down the scientific edifice. Sometimes the problems continue to accumulate and the paradigm reaches a point where a "crisis" arises. A scientific revolution may ensue - replacing the old science with an entirely new and different scientific paradigm. Scientific paradigms are "incommensurable" - fundamentally incomparable and incoherent with each other. Kuhn regards transition from one paradigm to another as like a "gestalt switch" or "religious conversion". [Chalmers, 2013, pp 100-101, et seq]

Imre Lakatos, introduced further nuances in the understanding of science. Uncomfortable with the relativist implications of Kuhn's schema, but recognising the limitations of Popper's framework, Lakatos understood science to be undertaken through a multiplicity of "research programs". A research program is comprised of a theoretical core supported by a belt of subsidiary hypotheses. The core cannot be permitted to be falsified as it represents an ideological commitment shared by researchers in the program (its "negative heuristic"). The belt serves to protect the core by being open to revision as needed to support the core. A research program can be described as progressive if it is leading to the discovery of new phenomena or degenerating if it is not. A research program is accompanied by a "positive heuristic" which provides guidance to researchers on the kinds of problems and methods which fall within the bounds of the research program. Lakatos' programs transition from one to another given their progressive or degenerative character.[Chalmers, 2013, pp 21 et seq]

The interested reader is referred to Chalmer's work, as the above is an incomplete account. Also, it is useful to emphasise that the purpose of this brief review is not to question the value of scientific knowledge or its success in generating increasingly accurate and practically useful descriptions of reality. Rather it goes to opening a more nuanced understanding of how science operates as a form of knowledge, in the context of the domain with which we are concerned. The review of the dominant legal theories carried out earlier in this chapter suggests that the legal domain is still

heavily influenced by the effects of rather simplistic notions of science. In the second edition of his work, Chalmers refers to the problem of misapplications of "science" in other domains.[54] It is a relevant cautionary tale in a multidisciplinary study seeking to explore law primarily from a computational perspective.

> " .... there is no timeless and universal conception of science or scientific method ... We cannot legitimately defend or reject items of knowledge because they do or do not conform to some ready-made criterion of scientificity. The going is tougher than that."[Chalmers, 1982, p 169]

So what are we to make of "doing" law in the multidisciplinary ways that have been explored above? What are the implications of "law as ..." insights? A new "scientific" theory of law is neither asserted, nor justifiable.

Legal positivism, assumes the external objectivity of law. It assumes that a value free observer or user of the law can rationally determine the nature of law without changing that legal character. In science such certainty no longer holds. The observer or measurement problem has been made famous by the quantum double slit experiment. The most widely accepted interpretation of the experiment leads to the following conclusion proposed by Neil Bohrs. *"The procedure of measurement has an essential influence on the conditions on which the very definition of the physical quantities in question rests."* Bohrs related this to the subjectivity suggested by special relativity. *"The theory of relativity reminds us of the subjective [observer dependent] character of all physical phenomena, a character which depends essentially upon the state of motion of the observer."* In lay terms, how we choose to observe the world, determines what we see in it. The experimental setup gives us equally correct but contradictory answers.[Krips, 2013]

Of course we are not here dealing with quantum mechanics. Nonetheless the example inspires an idea as to how to employ the "law as ..." approach, in ways that are helpful to our understanding of law but which avoid misleading simplifications. The insights derived from the "law as ..." approach can be conceived to be a kind of experimental design. They do not provide a single reductive and complete "truth" about law to be championed in contrast to other theories. Rather, each "law as ..." account can be considered a thought experiment in which law is observed using a different experimental setup. Rather than implying irresolvable contradiction in the diverse answers provided by each thought experiment - law wears the guise we constrain it to, according to the questions we ask of it. To obtain a fuller understanding of the nature of law - to more completely illuminate its nature - a diversity of questions

---

[54]Unfortunately, this reflection on implications for other fields seems to have disappeared from the current edition.

need to be asked and the insights of the resulting answers combined into a fuller understanding of the phenomenon of law. That law is about "rules" is undeniable - and valuable to know. Moreover positivist insights about rules have advanced our understanding of them. The problem is not, primarily the answers provided, rather it is the answers that have been ruled out. If we want to know about the law, rules are a very small part of a much more complex and interesting story. Law is also about communication, data, design, networks and complex adaptive systems. The "law as ..." approach offers an open agenda for investigation.

Further, the above has focussed on the question of what law "is". Thus far the discussion has avoided explicitly asking whether the foregoing provides any insights as to what the law "ought to be". Smits suggests that any interdisciplinary study of the law cannot avoid this normative question. He suggests that interdisciplinary studies of law, whether admittedly or not engage with law in this way.[Smits, 2014] Whether or not this is true in general, the foregoing does raise normative questions. To speak of 'better' design of law, for example, is to speak of how the law 'ought to be'. The case is similar when considering 'enhancing' communication of law. Further such goals carry implications as to who the law is for, or who it ought to be for. Thus, above, we cited the anti-democratic implications of law designed for lawyers alone. Smith suggests that we ought not to shy away from these normative implications. Rather he sees this as a contribution that interdisciplinary approaches to law can and should make to its study.[Smits, 2014]

## 2.2 Theoretical Frames

Above, we have investigated the nature of law from a multidisciplinary perspective. In that process, a theoretical background to the research of this thesis has been implicitly exposed. We now turn to that theoretical context, explicitly. In other words - we ask what are the theoretical frames of this research? As we have already seen, its multidisciplinary character compels a description of the research from multiple viewpoints. Legal informatics is an obvious candidate. Others are also relevant: access to law; visual communication of law; legal design and readability of law. Each of captures an aspect of the research.

1 This research is in the mode of legal informatics: it applies computational technologies to the legal field - more specifically for the enhancement of the communication of law. It thus falls comfortably within the field of legal informatics.

2 Enhancing the communication of law is a goal within the scope of access to law. It is thus concerned with the body of work and research carried on by the Free

Access to Law Movement.

3 Readability of law is one mode in which the communication of law can be enhanced and this research thus addresses the readability of law.

4 Visualisation of law is a way of enhancing communication of law which is both novel in its own right and is a mode for which computational technologies offer particular advantages. From this perspective, the research is concerned with visual communication.

5 Finally, this research engages the modality of legal design and applies design paradigms to enhancing the communication of law.

### 2.2.1   Legal Informatics

Above, we have considered the nature of law as data, information and knowledge. In this section, we focus on legal informatics as a discipline and theoretical framework for the investigation of law. Legal Informatics, as described below, arises in the context of the application of computers in the law. It is a still emerging body of scholarship concerned with studying information and computational technologies as applied to law.

The discipline has a brief history. It begins in the mid-1960s when mathematicians were considering the application of their discipline to law.[Paliwala, 2010]

Hinson reviews the literature for definitions of legal informatics. The definitions are various, highlighting different dimensions. Erdelez and O'Hare emphasise "application of informatics ... within the legal environment". Lucchi emphasises application of informatics "in the field of law" including "all uses of computers in law". For Sorkin, legal informatics is "the study of legal information systems and their use". Matthijssen emphasises the study of how "the work of legal practitioners can be supported through technology". Hinson himself defines legal informatics as "the study of information, its technology, and its implications and impacts in the field of law."[Hinson, 2005] Seipel provides a similar definition but also notes, helpfully, the interdisciplinary character of legal informatics. It is a "law and ..." field. It is concerned with both information science and the law.[Seipel, 2004]

Further useful insight into legal informatics is obtained by considering how the subject is taught. At Stanford University, Legal Informatics is taught with a focus on the application and impacts of information technology in the legal field (as suggested by the Hinson definition):

*"This class offers an overview of how technology is used in today's legal practice*

*and how it will be changing the landscape of the legal profession and the law more*
*broadly in the foreseeable future."*

Discrete modules address legal document management (search, e-discovery, specialized databases), legal infrastructure (systems for managing legal information including client and business information) and computational law ("expert legal systems", "computable contracts" and unauthorised practice of law).[55]

Two of the leading international conferences in the field of legal informatics are the biennial International Conference on Artificial Intelligence and the Law and the annual Jurix conference (based in Europe). ICAIL emphasises the development of computational methods for legal reasoning, application of AI to the law, discovery, machine learning and data mining and formal models of norms and normative systems.[56] The Jurix conference also emphasises artificial intelligence, considering both a theoretical and practical dimension. Theory concerns itself with models (e.g. of legal knowledge), representations (e.g. languages), methods and algorithms in application to norms. Practical application canvasses technologies for expressing legal semantics, for large scale analysis of legal information, for validation of legal knowledge systems, social simulations and information retrieval or management of legal systems. Practical applications were also considered across the legal domain (judiciary, legal practice, police, business, data protection, and public administration among examples).[57]

In 2010 a special issue on the history of legal informatics was published in the European Journal of Law and Technology. The articles published in the journal also illustrate the scope of legal informatics. The global movement for free access to law through online open access publication of legal materials appears in the issue.[Greenleaf, 2010] Developments in court room information technology is considered.[Martin, 2010] An article looks at the rise and then fall of expert systems, which made an early appearance in legal informatics.[Leith, 2010] Paliwala's introduction to the issue is somewhat down beat. Susskind is cited as negative about the uptake of technology by the profession. Martin is critical of limited court adoption of technology. De Mulder stated that jurimetrics, empirical study of law has 'never really come into being'. Expert systems failed because those who promoted them failed to understand legal culture. Having substantially lowered expectations, Paliwala then reverses direction noting achievements such as the dramatic transformation of 'the

---

[55] Legal Informatics Class - Spring 2014, Law 729/CS 204 http://codex.stanford.edu/legal-informatics-class/ accessed 27 August 2015.

[56] ICAIL 2015 Call for Papers http://sites.sandiego.edu/icail/call-for-papers/ accessed 27 August 2015.

[57] Jurix 1015 Call for Papers http://jurix.nl/jurix-2015-call-for-papers/ accessed 27 August 2015.

nature of legal publishing and the life of the law itself'; the 'worldwide flourishing of electronic legal information'; and the emergence of 'a worldwide culture of free access to legal information'.[Paliwala, 2010] We see a field which is still emerging and uncertain as to its boundaries. It is most confident and certain at its core - the application of computer technology to law.

The evolution of the field can also be illustrated by reference to the evolution of the ICAIL conference which is summarised by the review of its first 25 years in [Bench-Capon et al., 2012]. The initial conference included themes such as expert systems, case-based systems, and rule-based argumentation. The 1989 conference extended research on expert systems looking at various approaches to such systems. 1991 again saw legal expert systems as a strong area of focus. 1993 saw a broadening of focus with the introduction of game theory, heuristic search and neural networks. 1995 introduced new themes such as ontologies as well as continuing established areas of research. Jumping forward to 2011, the conference had extended its areas of investigation to include applications such as Bayesian reasoning, risk analysis and agent software engineering.[Bench-Capon et al., 2012]

In overview we see a broad field which at its heart is about application of computer technologies in the legal field.

### 2.2.2 Access to Law

The World Wide Web was developed over 1990/1991.[Connolly and Cailliau, 2000] Its essential innovations were the combination of a number of technologies: hyptertext (an already existing concept of linking documents by embedded links); the HTTP protocol for access to resources provided by servers on the web; a uniform resource locator for each resource; a web browser (a client program for accessing resources on the web) and HTML (a mark up language for text documents adapted from SGML). [Berners-Lee, 1989; Berners-Lee et al., 2006; Berners-Lee and Cailliau, 1990; Berners-Lee, 1991a,b; Foundation, 1991]

Those who have grown up in the era of smartphones may not appreciate the wonder of the ability to access, instantaneously, knowledge from anywhere in the world. In the 1990s anyone who opened a web browser for the first time could not have failed to appreciate the transformative breakthrough the web represented.

Later the primary inventor of the world wide web, Tim Berners-Lee was to establish the World Wide Web Foundation for the purposes of promoting free access to the web, accessibility and access prioritising people over organisational interests.[58]

---

[58]World Wide Web Foundation About Page http://webfoundation.org/about/ accessed 14 September 2015

These goals were noted to have arisen from certain ideas and values that emerged from the web community itself. These values included: decentralisation of control; non-discrimination in access to data; bottom-up and open design; universal access to publication; consensus on standards. The foundation notes that on the basis of these concepts movements such as open data; open government; scientific open access and free culture emerged.[Foundation, 1991].

Lawyers were among those who saw the potential of the web. The Legal Information institute emerged in 1992 and by 1994 was providing access to United States legal materials on the web. They were quickly followed by similar movements in Canada and Australia, and later in other jurisdictions. These national developments were brought together in the Free Access to Law Movement (FALM) which was established in 2002.[St Amant, 2007, p375] FALM, comprised of 50 member organizations around the world, maintains a secretariat and hosts the annual Law via the Internet Conference.[59] Since 2013 the movement has maintained the Journal of Open Access to Law. The goal of the journal is to promote research on open access to law.[60] In 2002 FALM adopted the Montreal Declaration on Open Access to Law, a founding document which has been subsequently revised, most recently in 2012. The Declaration includes the following statements:

- *"Public legal information from all countries and international institutions is part of the common heritage of humanity. Maximising access to this information promotes justice and the rule of law;*

- *Public legal information is digital common property and should be accessible to all on a non-profit basis and free of charge;*

- *Organisations such as legal information institutes have the right to publish public legal information and the government bodies that create or control that information should provide access to it so that it can be published by other parties."*[61]

These statements provide the most widely accepted definition of what open access to law means. It may be noted that the statement draws on concepts from property law and international law. It prioritises access 'free of charge' and calls on government bodies to 'provide access'. The concept of the common heritage of mankind finds its early expressions in international treaties on law of the sea and outerspace.[Wolfrum, 1983] The concept is invoked when resources "outside the limits of national jurisdiction" are at stake.[Wolfrum, 1983] It is natural that this idea

---

[59]The Free Access to Law Movement http://www.fatlm.org/ accessed 14 September 2015.

[60]Journal of Open Access to Law https://ojs.law.cornell.edu/index.php/joal/index accessed 14 September 2015.

[61]FALM Declaration http://www.fatlm.org/declaration/ accessed 14 September 2015

should suggest itself as applicable to "cyberspace". The properatarian focus of the Declaration also in part is a function of the 'problem' FALM sought to solve: namely the privatisation of public legal information based on copyright ownership. Thus, as stated by Greenleaf, in respect of Australian developments:

> *"The international development of public legal information servers is part of the more general movement to create publicly available (or 'free to air') resources on the Internet, similar in some respects to the creation of public libraries in the nineteenth century. The Internet is fast becoming home to commercial providers of information, and effective means of charging for even occasional uses of resources are being developed. The countervailing movement, of which AustLII is a part, aims to ensure that some part of cyberspace is public space, where no one is denied use of resources because of financial considerations."*[Greenleaf et al., 1995]

To these central concerns may be added a number of additional understandings of open access, some of which have informed this research.

Firstly a practical focus on "access" concerns itself with the user's experience of access to law. In this respect, the New Zealand Parliamentary Counsel's Office identified three factors which bear on accessibility: that the law is available online; that it is navigable; and that it is understandable.[NZ, 2008]

Secondly, in addition to property based arguments other arguments for open access to law include:

- that the rule of law depends on law being available and understandable to its audience;

- that it is unfair to require citizens to obey law that that they are unable to understand;

- that poorly communicated law is ineffective as a legislative exercise;

- that economic efficiency is enhanced by effective access;

- that the law should be written for its audience (now including the public at large);

- that there is or should be a right to open access to law; and

- that it is implicit in democracies that laws should be accessible to citizens.[62]

---

[62]See discussion in Section 2.1 of [Curtotti et al., 2015c], (Appendix A2).

The research reported in this thesis seeks to contribute to open access to law. It focusses on the understandability and communication of law. Three of the papers forming the body of this research were presented at the Law via the Internet Conference, and two are published in the Journal of Open Access to Law.

### 2.2.3 Readability of Law

A further frame of reference for the research reported in this thesis is the body of knowledge concerned with readability of language and the readability of law specifically.

Readability research has particularly focussed on measuring the readability of text for particular audiences through the development of readability metrics. Historically, most readability work was undertaken for the purpose of ranking reading material for educational purposes.[DuBay, 2004] Measures developed in that context were applied to the more general task of reviewing the readability of adult reading materials - although not designed for that purpose.[Woods et al., 1998]

More recently, a flood of research has applied natural language processing and machine learning to the prediction of text readability.[Collins-Thompson, 2014] Research on the readability of legislation has also been undertaken, although in a much smaller volume than readability research in general. Research has applied readability metrics and cloze tests to measure the readability of legislative materials. However concern about the poor readability of legal material has been expressed through the plain language movement which has progressively transformed the writing of the law. A concern for readability of legal materials has a long history - stretching back to the replacement of 'legal' Latin and 'legal' French in medieval England. The 19th century saw a particularly fertile period of reform that essentially set the structure and form of legislative expression that is still used today in many common law countries.[Curtotti and McCreath, 2012; Curtotti et al., 2015c]

### 2.2.4 Visual Communication of Law

Visual communication has been a further frame of reference for the research reported in this thesis. Text dominates the communication of law. Visual communication, however, engages different cognitive processes than oral communication and pure text. An example of visual communication of law is provided above in Section 2.1.3, in Governor Davey's proclamation. Figure 2.12 provides a further example of visualization of law.[63] The visual focusses on a key issue of interest to the addressed

---

[63]U.S. Student Aid Infographic https://studentaid.ed.gov/sa/eligibility accessed 17 September 2015. Image in public domain.

audience (eligibility for benefits) and sets out in text and visuals conditions for eligibility. Conjunctive conditions of eligibility ('ands') are illustrated in a serial pathway while disjunctive conditions ('or') are illustrated as branching alternative pathways.

Visual communication engages a multidisciplinary field.[Smith et al., 2005; Moriarty and Kenney, 1995] The Handbook of Visual Communication, for example, canvasses visual communication theory from a variety of viewpoints including aesthetics (beauty); perception theory, representation theory, visual rhetoric, cognition theory and visual semiotics.[Smith et al., 2005] As commented in its preface: *"we humans had to learn how to walk, talk and read, we never had to learn how to see ... [thus] we are less cognizant of the many processes that contribute to what we see."*[Smith et al., 2005, p ix]

The visual communication of law fits more specifically within the domains of information visualization or knowledge visualization. The former has largely developed in the context of the presentation of data given the previously unavailable processing capacities of computers. However, there is a lack of a coherent widely accepted theoretical framework for information visualization.[64] Emphasis may be on science or on software engineering. Cognitive science perspectives of information theoretic perspectives may be prioritized.[Anderson et al., 2010; Grammel and Storey, 2010; Swienty and Takatsuka, 2010; Yi, 2010; Chen and Jaenicke, 2010] Information visualization is often understood to mean "computer assisted use of visual processing to gain understanding".[Curtotti and McCreath, 2012] (See [Burkhard, 2004; Card and Mackinlay, 1997; Lengler and Eppler, 2007]) Knowledge visualization is slightly broader in conception and not confined to the computational context and emphasises the communication of knowledge rather than information. Thus, Burkhard states:

> *Knowledge visualization examines the use of visual representations to improve the transfer of knowledge between at least two persons or group of persons.*[Burkhard, 2004]

The distinction between information visualization and knowledge visualization also engages a discussion of the data –> information –> knowledge triad discussed above in section 2.1.5. We saw in Zins' analysis of theoretical frames for the communication of D-I-K the existence of subjective and objective interpretations of these entities with corresponding emphasis on an agent's subjective experience or on the objective manifestations of D-I-K in symbols representing stimuli, knowledge or meaning.[Zins, 2007]

---

[64]See discussion in Section 4 of *Enhancing the Visualization of Law* [Curtotti and McCreath, 2012] (Appendix A.6)

Rather than attempting a review of all these theoretical frames, semiotics is focussed on below, as it provides a methodology for review of the visualization of law discussed in this thesis.

Semiotics studies signs. A sign is "anything that stands for something else ... a sign stands for an object or concept". Messages are communications composed of signs. Understanding depends on the receiver's knowledge of the code - i.e. the meaning attached to the signs. Two early founders of semiotics were Ferdinand de Saussure (a linguist) and Charles S. Pierce who was concerned with epistemology. Saussure modelled signs as a diad of signifier and thing signified. Pierce added a subjective element making a triad of signifier, signified and "interpretant" (the subjective mental model created in the recipient of the message). Pierce distinguished a hierarchy of three different types of signs: *iconic* - essentially a picture of the thing represented); *indexical* - the image points to the thing represented (e.g. a raindrop for water); and *symbolic* the image symbolises the thing or concept by convention (for example letters by convention represented certain sounds). These foundation concepts are extended in various ways by scholars of semiotics. Signs may be motivated or unmotivated. Motivated signs have artificially imposed meaning - whereas unmotivated signs inherently express the thing signified. Words are generally inherently motivated - i.e. their sound has no natural connection to what they represent. Visual symbols may be unmotivated (as in art). Signs have denotations (a literal or immediate meaning) and connotations (implied meanings). Moreover signs have a plurality of meanings and from a network of chained meanings of signifier - signified - signifier (a potentially "unlimited process of interpretation"). The interpretation of signs involves a process of reasoning described as abduction. (See discussion in [Moriarty, 2005])

> *In contrast to inductive (reasoning to) and deduction (reasoning from) logic, abduction is an inferential process that fashions conjectures based on "clues" that are available or conditions that are known. ... In order to accumulate clues, the abductive process begins with observation, the bits and bytes of perception. It's similar to the way a doctor accumulates symptoms until he or she arrives at a diagnosis. Peirce described the formation of an abductive hypothesis as "act of insight," the idea coming "like a flash" – the proverbial light bulb. In a more formal statement, abductive reasoning assembles the observations and attributes a variety of characteristics or conditions to a subject (the conjecture process) until a match is made and a conclusion can be stated. Beyond the linear forms of deductive and inductive logic, abductive reasoning more closely resembles massive parallel processing by computers, one that is not at all like language processing.*

> *The processing, ... is close to the nonlinear pattern used in perceptual processing, as Barry (1997) has explained in her book, Visual Intelligence.*[Moriarty, 2005]

Another concept relevant to semiotics is that of "code". A code is a broad concept that encompasses understood frameworks of communication and culture. Thus grammar is a code for the structuring of language. Conventional representations in movies provide codes (for example infinitely repeated sequences such as car chases). Myths have codes embedded within them. (See discussion in [Moriarty, 2005].)

Visual semiotics may be employed to interpret the meaning conveyed by visual signs (e.g. images, scenes). This may be achieved by an analysis of images by a semiotician. Studies involving the creators of images/signs may also enable enhanced understanding of intended meaning. The semiotician may carry out empirical studies in which the received meaning of an audience is analysed. (See [Moriarty and Sayre, 2005; Dunleavy, 2005].)

In *Enhancing the Visualization of Law*,[Curtotti and McCreath, 2012][65] a review is undertaken of the current state of visualization of law, particularly in online environments. This review, unsurprisingly, confirms the largely textually based nature of visualization, but also notes the more visually oriented aspects. At the most basic, visualizations are pure text. From that base a variety of visualizations are identified:

- the use of layout, heading and numbering features developed in 19th century United Kingdom legislation;

- the conceptualisation and presentation of sections within legislation as hyperlinked documents;

- the use of document icons, in which law is conceived as downloadable documents reflecting the print edition statutes on which they are modelled;

- the 'enrichment' of legislation by providing information filtering by time, geography or theme, clickable footnoting and multiple versions; and

- the use of colour and font and contextual links relevant to legal information.

The use of images in a more historical context is illustrated in traffic regulations, which readily deploy image to convey the meaning of particular traffic rules. More experimental or novel work sets out to consciously use images to represent the *meaning* of law.[66]

---

[65]See Section 2 of Appendix A.6
[66]See Section 2.5 of [Curtotti and McCreath, 2012] (Appendix A.6)

The use of visual communication in the law has attracted increasing interest in recent years.  Rosman provides an introduction to the use of visuals in the law - particularly in the context of legal argumentation.  He illustrates how visual presentations such as timelines, flowcharts, tables, decision trees can more readily communicate key facts and issues in legal cases.[Rosman, 2013] Porter also reviews the use of visuals in legal materials noting the anti-image culture of the law.

> *"Tradition governs every aspect of a court opinion ... And according to that tradition – which in large part predates the camera, never mind the computer – images have a peripheral or, more typically, nonexistent role.  Law has been trapped in a stylistic straitjacket.  The Internet has revolutionized media and communications, replacing text with a dizzying array of multi-media graphics and images. ... But those innovations have barely trickled into the black-and-white world of written law.  Legal treatises continue to evoke Blackstone and Kent; most legal casebooks are facsimiles of Langdell's; and legal journals resemble the Harvard Law Review circa 1887.  None of these influential forms of disseminating the law has embraced – or even nodded to – modern, image-saturated communication norms.  Lawyers and courts routinely confront visual questions, ... But courts, scholars, and practitioners analyze such image-centered disputes within ... a framework in which the alleged objectivity of text literally papers over the emotion-laden visual subjects in dispute.  Images are associated with emotion and irrationality.  Written law resists that irrationality ... "*[Porter, 2014]

However, Porter argues that like other print bound media that have become saturated with multimedia content, the law is facing the same future, which is already being embraced by the young and innovative in the law.  In that context Porter identifies visual literacy as a gap in the education of lawyers.  As visual communication enters the court and the law, lawyers are unprepared for the implications.  In addition to describing the use of visual communication in the law, Porter argues the need to develop principles for the conduct of visual argumentation.  She provides a number of examples of questionable use of images in court proceedings.[Porter, 2014] Interestingly, she makes no reference to semiotics, which would appear to be centrally relevant to the need to develop "visual literacy" in the legal profession.  This is despite the existence of a school of legal semiotics.[67]

The availability of computers, as pointed out by Porter and others is fundamental to the new potential for the visual in law.  The American Bar Association Journal in 2014 thus ran an article under the title *Visual law services are worth a thousand*

---

[67]See for example the International Journal of Semiotics of Law http://www.springer.com/law/journal/11196

*words–and big money.* The article reviews computationally mediated visualizations of law including the work of Ravel (with its visualization of citation networks for legal research purposes), and Margaret Hagan's work teaching visual legal and legal design.[68] Yale University runs a "Visual Law Project" that aims to teach law students "visual advocacy" by providing education in film making.[69] The Interdisciplinary Centre for Law and ICT runs a Visual Law Lab that aims to "aims to provide a forum for reflection and experimentation ... on how visualisation of legal norms ... could contribute to conveying legal information in a more meaningful way - making it easily accessible and understandable for the intended audience - and hence, to increasing the efficiency of existing laws and regulations in steering human behaviour."[70]

One of the pioneers of visual law in the European context is Colette Brunschwig. She also notes the dominance of text in the legal world, and examines the question of whether the legal world is already making a "visual turn". She concludes that there is already a "powerful trend" towards visual law. She thus argues that the issue now is to explore the implications of these new developments. *"There is a strong need for a legal discipline capable of exploring all visual legal communication practices."*[Brunschwig, 2014] Brunschwig's comments suggest a discipline that is still in its early stages, as might be expected of a disciplinary connection that impinges on a traditionally textual world through relatively recent accessibility of image that computational power makes available.

Johansen and Robbins seek to provide guidance to legal writers on the use of visuals. They note, in a context in which professionals have or are shifting to electronic devices and away from paper, the wide ranging scholarship validating the benefits of inclusion of visual communication in legal writing. However: *"[t]he questions of what, why, how, and when to use visuals in legal writing remain open."* They propose the greater use of visuals in legal analysis, in addition to the already existing use of visuals to better communicate facts in issue. Figure 2.13 illustrates Johansen and Robbins' analysis of use of legal visuals in legal argumentation. The four quadrants of Figure 2.13 summarise their recommendations on the 'what, why, how and when' of legal visuals.[Johansen and Robbins, 2015] Under 'what', Johansen and Robbins propose a taxonomy of legal visuals. Interestingly, their taxonomy does not build on the taxonomy of signs or the scholarship of visual communication. That scholarship might have been usefully referred to in classifying visuals for the purposes of legal

---

[68]Visual law services are worth a thousand words - and big money http://www.abajournal.com/magazine/article/visual_law_services_are_worth_a_thousand_words_and_big_money

[69]Yale Visual Law Project http://yalevisuallawproject.org/ downloaded 17 September 2015.

[70]University of Lueven Interdisciplinary Centre for Law and ICT, Visual Law Lab https://www.law.kuleuven.be/icri/en/news/item/visual-law-lab downloaded 17 September 2015.

rhetoric. Nonetheless, their paper presses the boundaries of the field and is further evidence of the increasing interest and engagement with visual communication in the law.

## 2.2.5   Legal Design

The topic of visual communication of law brings us naturally to the question of Legal Design, as frequently "legal designers" use visuals as part of their methodology. We have already reviewed design in the context of examination of law as designed artefact (see Section 2.1.8) and also in the context of the implications for law of its characteristics as complex adaptive system. (See Section 2.1.9) That discussion highlighted such design principles as user centredness and addressing a hierarchy of user needs from functionality to user experience. It also highlighted the difficulty of 'designing' a complex adaptive artefact that is itself evolving and reacting to its environment, which in turn is reacting to it.

Legal design also represents a recent movement that focusses on the application of design principles and methodologies to the law and legal communication. Visualization is a prominent tool in the legal design context.

The Legal Design Initiative runs out of the Stanford University Institute of Design develops software and undertakes teaching and research. It provides a visual law library, showcasing examples of visual law. They also provide a visual design toolbox for lawyers which overviews principles of design as well as a relevant reading list.[71]

Legal Design Jams are part of a growing movement focussing on the 'design' of law. One site describes Legal Design Jams as follows:

> *"A Legal Design Jam brings together a group of motivated individuals from different fields (e.g. designers, lawyers, policy-makers, coders, innovators, business people) and, together, give an extreme user-centric makeover to a legal document. The idea is borrowed from hackathons and service jams, and seeks to engage people to rethink and innovate the very concept of what a legal document should be, look and feel. ... Citizens, users, consumers, tenants all have the right to access legal information that is clear and well-communicated. ... "*[72]

The Legal Design Jam uses design concepts such as 'user-centredness'. It notably cites a rights based argument for legal access. It prioritizes the design of documents using visualization, simplification of language, multidisciplinary approaches and ap-

---

[71]Legal Design Initiative http://www.legaltechdesign.com accessed on 16 September 2015. See also Open Law Lab http://www.openlawlab.com by Margaret Hagan of the Legal Design Initiative

[72]Legal Design Jam About http://legaldesignjam.com/about/ downloaded 17 September 2015.

plying 'information design'.[73] Design Jams have been held across Europe and the Americas.[74] As an example, the Legal Design Jam held at the Simplification Centre of the Aegean University in 2013[75] worked on re-designing the United Nations Convention on Contracts for the International Sale of Goods (CISG). Figure 2.14 provides an extract from the prototype re-design produced by the Jam. The figure shows the principles for application of the CISG.[76]

## 2.3   Conclusions

This chapter has provided a theoretical investigation and framing for the work reported in this thesis. It has looked outwards from the reported work. It asked the question: what are the implications for the nature of law which arise from its multi-disciplinary investigations? A 'law as' framework was adopted to found an enriched description of law. Law can legitimately be described as document, rule, communication, network, computation, designed artefact and complex adaptive system. These facets of law are not mutually exclusive. Taken together they provide a fuller account of the nature of law and add usefully to the body of jurisprudence that already exists concerning the law. Section 2.2 returns to a more conventional starting point for a thesis: its theoretical background. As the section illustrates (complementing investigations on the nature of law) - the theoretical frames for this research are multidisciplinary. The law is of course one frame of reference. Other relevant frames are legal informatics, access to law, readability of law, visual communication of law an legal design. In the chapters that follow the discourse turns to a description of the data, experiments and results of the reported work, to reflections on the body of the research and to the contributions made by this thesis and the reported work.

---

[73]Ibid.

[74]Legal Design Jams - Past Jams http://legaldesignjam.com/about/past-jams/ downloaded 17 September 2015

[75]Legal Design Jam Report http://legaldesignjam.com/ldj-syros-report/ accessed 17 September 2015

[76]CISG Legal Design Jam Group @ Syros 2013 Visual CISG - A Prototype of Legal Information Design http://legaldesignjam.com/wp-content/uploads/2013/11/CISG_booklet.pdf downloaded 17 September 2015 (made available under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (CC BY-NC-SA 3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/))

Figure 2.12: Student eligibility for Federal Aid in the United States

Figure 2.13: A summary of Johansen and Robbins' analysis of the use of visual communication in legal argumentation



Figure 2.14: When does the CISG Apply?

# Data, Technologies, Experiments, Tools, Dissemination

The purpose of this chapter is to provide a description of the practical work undertaken during the research reported in this thesis. It describes the data collected and analysed during the research. Technologies and experimental methods are outlined. Tools used included programming languages, client server architectures and software for natural language processing and machine learning, all of which are summarised. This chapter concludes with a description of how the work reported in this thesis has been disseminated.

## 3.1 Data

### 3.1.1 Australian Corpus of Contract Language

The initial body of data studied during this research was a corpus of Australian Contract Language. The corpus was compiled from Australian contracts collected from the world wide web. After data cleaning the corpus comprised around 250 contracts with an overall corpus size of approximately 1,000,000 words. A sub-corpus of 30 contracts extracted from the Australian Contract Corpus was the subject of a study of classification of contract text reported in [Curtotti and McCreath, 2010]. A corpus study on the corpus as a whole was carried out in [Curtotti and McCreath, 2011].[1] Also sub-corpora of 10, 20 and 30 contracts were extracted from this corpus and used in the analysis of legal definitions for the purpose of visualizing definition networks.[Curtotti et al., 2013][2] The Australian Contract Corpus has been made available to researchers on request, and has been provided to a number of researchers who have requested access.

---

[1]Appendix A.7
[2]Appendix A.5

### 3.1.2   Australian Legislative Corpus

A corpus of Australian legislative texts was also created and used during the research. This corpus was compiled from 'popular' items of Australian legislation as listed on the official Australian government legislation website. The corpus was the subject of a comparative study with other English corpora for the purpose of assessing differences with other English genres. This work was reported in [Curtotti and McCreath, 2013].[3]

### 3.1.3   American Corpus of Regulatory English

A corpus of U.S. regulatory English was created from legislative sentences randomly selected from the United States Code and the U.S. Code of Federal Regulations (approximately 500 sentences each). The corpus is provided in three versions. This first version is a raw text version of the corpus. The second version is an enhanced marked up version tagged with context free grammar parts of speech tags. The third version is an enhanced marked up version tagged based on the Stanford Dependency Grammar. The corpus was used for readability testing in a large online study involving over 14900 respondents and over 43,000 responses reported in [Curtotti et al., 2015c] and [Curtotti et al., 2015b].[4] The corpus also includes labels for readability ranking based on the results of user testing carried out and reported in the same research. Already a researcher has requested access to the Corpus.

### 3.1.4   Non-Legal Corpora

In addition to these legal corpora, a number of non-legal corpora were used during the research. A corpus of graded reading materials was extracted from a website and used in the research reported in [Curtotti et al., 2015c] and [Curtotti et al., 2015b].[5] The Brown Corpus of American Written English [Francis and Kucera, 1964] was used extensively during the research as a comparator for 'standard' English ([Curtotti and McCreath, 2011, 2013; Curtotti et al., 2015c,b] Other non-legal corpora used during the research were: the Reuters Corpus, the ABC news corpus (rural and science reports), Emma by Jane Austen and Movie Review Corpora and were used in [Curtotti and McCreath, 2011]. Apart from the graded reading material, all corpora were available through the Natural Language Toolkit.[Bird et al., 2009]

---

[3]Appendix A.4
[4]Appendices A.2 and A.1.
[5]Downloaded from http://www.lextutor.ca/graded/. Now no longer available at the website.

### 3.1.5   Readability Experimental Data

In the later phases of research (reported in [Curtotti et al., 2015c,b]), online data was collected from the Legal Information Institute website, including crowd sourced user responses to readability testing of legislative and non-legislative sentences. The data comprised of:

- 12 months page visit data from Google Analytics for legislative pages on the Legal Information Institute website;

- c. 15,000 demographic data submissions provided by users of legislative pages on the Legal Information Institute website;

- c. 13,000 cloze testing results on test sentences drawn from prepared corpora provided by users during online crowdsourced testing;

- c. 23,500 Likert test results on the same test sentences, similarly crowdsourced; and

- c. 12,000 semantic differential test results on the same test sentences, also crowdsourced.

In addition to the above data, Google Analytics for the LII legislative pages were analysed and contributed to the results of the research.

### 3.1.6   Visualization of Legislation

In [Curtotti and McCreath, 2012][6] a study is undertaken of the visualization of law. The sources of data for this study included websites publishing law (particularly in English speaking jurisdictions), sites investigating and publishing novel visualizations of law, historical and legislative materials and the research literature addressing visualization of law.

### 3.1.7   Case Study in Automated Visualization of Contract Clauses and Reflective Research

In [Passera et al., 2014] the automation of visualization of contract clauses is reported. The research took business to business contracts as case studies focussing on provisions found in such contracts: e.g. termination provisions, liquidated damages provisions and payment clauses. The results of this research and the process of creating the visualizations then became input data for the study reported in [Curtotti

---

[6]Appendix A.6

et al., 2015a][7] which reflected on learning for interdisciplinary research arising from [Passera et al., 2014].

## 3.2   Technologies and Experimental Methods

### 3.2.1   Natural Language Processing

Natural language processing (NLP) uses general knowledge about natural language in computers to process text documents.[Feldman and Sanger, 2007] A broader definition is "any kind of computer manipulation of natural language" - from "counting words" to "understanding language".[Bird et al., 2009, Preface] It often involves a pipeline of transformations, a typical example of which is illustrated in Figure 3.1. As will be evident, every step of the process requires representation in a computer, thus text (a concatenated list of characters in computer memory), but may be enhanced - by tokenisation (identification of words, numerals, punctuation), parts of speech tagging, representation of grammatical sentences structure and in more advanced processes by information extraction (such as named entity extraction - e.g. identifying individuals and other specific entities mentioned in the text).[Feldman and Sanger, 2007, pp59 et seq.] NLP was used extensively during the research and provided a backbone tool for the bulk of the research. NLP techniques applied during the research included segmentation at letter, word and sentence level (e.g. for construction of n-grams), parts of speech tagging, chunking and syntactic parsing to produce either context free or dependency grammars.
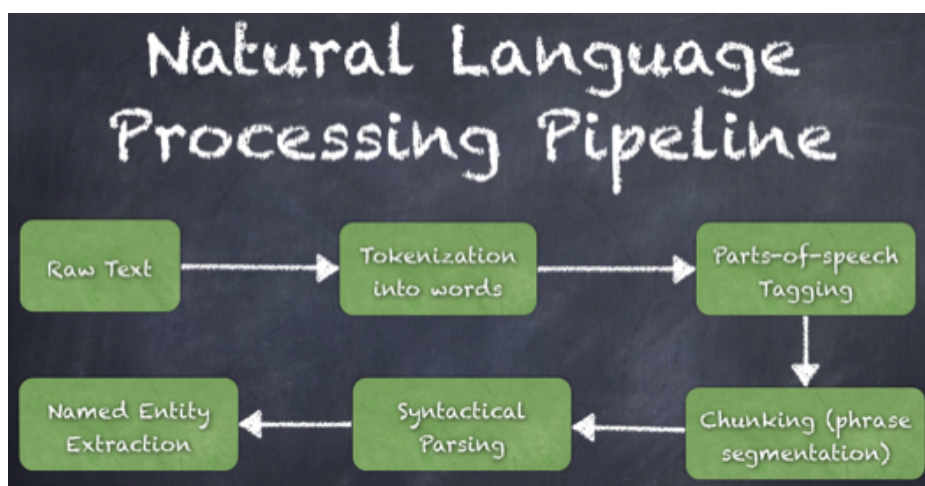


Figure 3.1: A Typical Natural Language Processing Pipeline

---

[7]Appendix A.3

### 3.2.2   Corpus Linguistics

Corpus linguistics is intimately connected to NLP. Corpus linguistics focusses on the data, as opposed to the processes. It is concerned with collections (corpora) of naturally occurring language. Such corpora form "the raw fuel of NLP". A corpus is "a body of machine-readable linguistic evidence". Corpora can be used for a variety of purposes. They may be used for the description and study of language. They may be used as test beds for the development of applications.[McEnery, 2005] As will be evident from the principal data investigated during this research, corpus linguistics forms a basic mode of the investigation which were carried out.

### 3.2.3   Graph Analysis

As discussed in Section 2.1.6, law forms a mathematical graph comprised of nodes and edges. The graph characteristics of law were studied and utilised during the research. In [Curtotti et al., 2013][8] graphs of relationships between clauses and defined terms in contracts are constructed. This required data representing these relationships to be first extracted using natural language processing and then represented as nodes and edges. The graph characteristics of the resulting network are then used as input to visualisations of the graph characteristics. In [Curtotti et al., 2015b], dependency grammar characteristics (which also constitute mathematical graphs) are extracted and used as input features for readability classification of legal sentences.

### 3.2.4   Machine Learning

Machine learning is described in a variety of ways. One definition is that that machine learning comprises computational algorithms that seek to "emulate human intelligence by learning from the surrounding environment". More specifically such algorithms are not "hard coded" but learn by "experience or repetition".[El Naqa and Murphy, 2015]. In more concrete terms common forms of machine learning are supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves learning to predict the label of an unseen data example, from experience gained from being provided with already annotated training examples. Unsupervised learning is more in the nature of pattern recognition: for example clustering data into naturally occurring categories. Reinforcement learning involves the algorithm adjusting its predictions based on the examples it encounters and rewards assigned for correct prediction. Supervised learning is also described as a function in which the algorithm must predict an output from a given set of intputs.[Russell

---

[8]Appendix A.5

and Norvig, 2003] Machine learning was used right from the beginning of the research reported in this thesis for the purpose of automatic segmentation of contract text according to legal purpose of given text.[Curtotti and McCreath, 2010] It was also used later in the research to study the characteristics of legislative language, and to investigate and develop indicators for readability of legislative texts.[Curtotti and McCreath, 2013; Curtotti et al., 2015b,c][9]

### 3.2.5 Information Visualization

Information visualization is a form of communication. (See discussion in in Section 2.2.4). During the research computational techniques were used to automatically visualise definition networks in legal contracts, time-lines for termination of contracts and payment and liquidated damages clauses for commercial contracts.
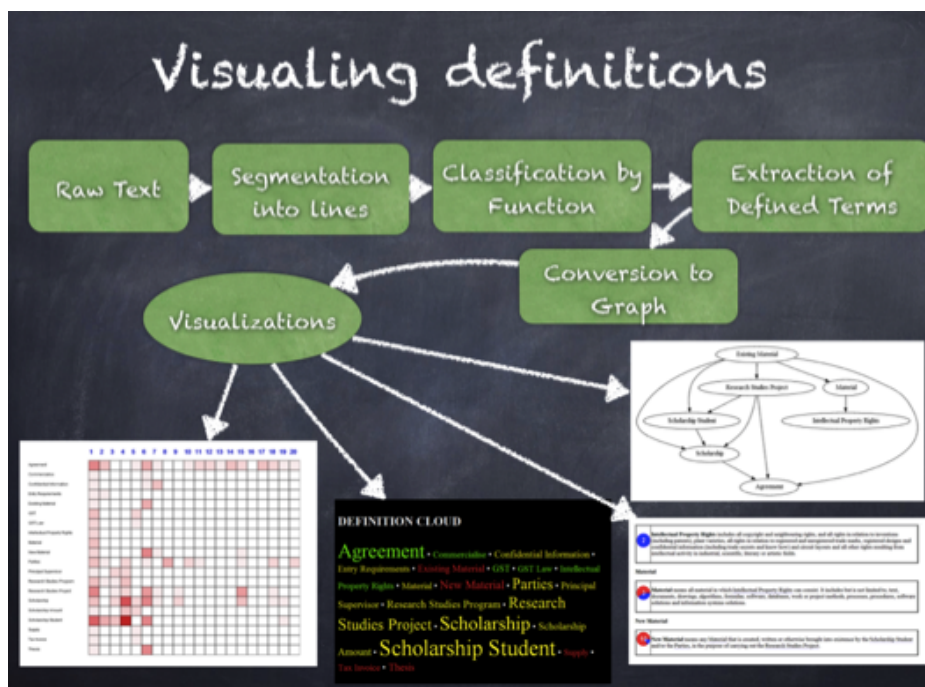


Figure 3.2: Processing Contract Text for Definition Network Visualization

Figure 3.2 illustrates the process for creation of definition visualizations. The process starts with the input of raw text, initially segments the text into lines, undertakes functional identification (including defining text), extracts defined terms and defining text, detects nodes and edges and generates a graph representation of the definition network within the contract. This data representation is then used to

---

[9]Appendices A.4, A.2 and A.1.

generate a number of alternative visualizations according to different potential use cases for the visualizations.[10] The visualizations also include hover and click based features which provide users with enriched information and allow the direct navigation of the definition network. This latter feature provides a prototype tool assisting contract drafters and readers.[11]

Figure 3.3 illustrates prototype visualizations of a termination clause from a legal contract. In this case, visualization starts with an abstract data model of the visualization. Variable data is collected from user interactions also provides the basis for for compiling alternative fixed texts versions of the clause, to accompany the visualization. The role of computational tools is to automate the production of a visualization of the data. This step would otherwise require the creative input of a human designer to achieve. Given the traditional aversion (and lack of training) of lawyers in visual communication, the automation of such visualization is potentially valuable to the creators of contracts.
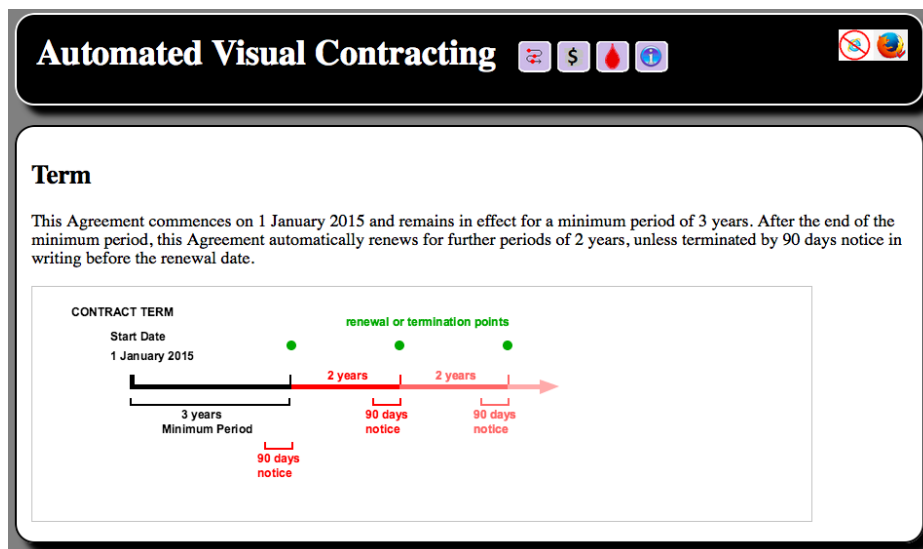


Figure 3.3: Visualization a Contract Termination Clause

A further example of such prototype visualizations is provided by Figure 3.4 which visualises weekly payment based on production levels. Each unit delivered receives payment of $100, however penalty and bonus provisions can reduce or increase overall weekly return and the impact is accurately reflected in the diagram.

---

[10] Visualizations reproduced from [Curtotti et al., 2013].

[11] See further [Curtotti and McCreath, 2012; Curtotti et al., 2013], appendices A.6 and A.5.
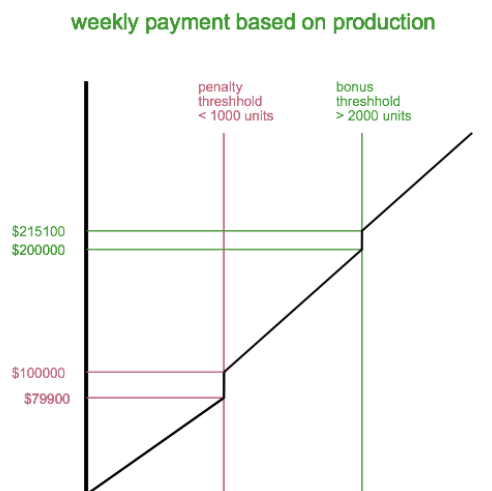
Figure 3.4: Payment by Production with Penalty and Bonus Provisions

### 3.2.6    Citizen science

The later stages of the research engaged crowdsourced techniques in the form of citizen science. Citizen science is defined by the Oxford English Dictionary as "scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions".[Curtotti et al., 2015c] In the context of this research users in an online environment were invited to participate in research by providing readability assessments of legal and non-legal sentences. The use of citizen science addressed one of the key challenges of undertaking machine learning - generating labelled data that can be used as training data for supervised learning.[Curtotti et al., 2015c,b]

### 3.2.7    Measuring readability

One of the key goals of the research was to learn to rank or classify legal sentences by reading difficulty. This required the application of methods of measuring readability. Classical readability metrics provided a background to the research.[DuBay, 2004] They were however inadequate to the task of readability of legislative language - not having been designed for that purpose. Accordingly user based measurements were investigated as appropriate measures - these measures were Likert questionnaires (a user selects options on the basis of their level of agreement with a statement), cloze tests (the user guesses missing words) and semantic differentials (a user rates a test stimulus against a number of qualities - for instance readability). Aggregate mea-

sures based on these measures provided input labels for machine learning. Aggregation was undertaken using principal components analysis.[Curtotti and McCreath, 2013; Curtotti et al., 2015c,b].

## 3.3  Software Tools

All stages of the research reported in this thesis involved software development or the utilisation of existing software packages. Software which was significantly used during the research included: the Natural Language Toolkit (for natural language processing),[Bird et al., 2009] the R-Statistical Package (primarily for data analysis),[R-Core-Team et al., 2012] the Weka Data Mining Software (for machine learning),[Hall et al., 2009] and the Stanford Dependency Parser (for natural language processing).[De Marneffe and Manning, 2008]

Programming languages used during the research included Python (for online platform development and text processing), Javascript and PHP (for interactive web functionality and visualization generation) and Java (for accessing the Stanford Dependency Parser).

## 3.4  Dissemination

The primary vehicles for dissemination of the research have been conference presentation and published papers. Generally, papers have been made available on an open access basis. In addition to these usual vehicles, the research has also been disseminated as follows:

- A "Readability Research Platform" was designed and made available to researchers via an ANU hosted web server. The platform provides an interface for undertaking readability research. Tools are provided for generating readability metrics, for parts of speech and phrase tagging. The website also provided get and post facilities for automatically submitting text to the platform for the purpose of extracting the outputs provided by the website.[12] This platform was also used for parts of speech tagging and chunking for the research reported in [Curtotti et al., 2015c] and [Curtotti et al., 2015b].[13]

- The prototype tool for automated visualization of clauses from legal contracts has also been made available via an ANU webpage.[14]

---

[12]The Readability Research Platform http://buttle.anu.edu.au/readability/ last accessed 25 September 2015

[13]Appendices A.2 and A.1.

[14]http://cs.anu.edu.au/people/Michael.Curtotti

- The research was introduced in an informal article published in Voxpopulii, an online blog of the Legal Information Institute.[15]

- A submission, based on the results of the research, was written and provided to the United Kingdom Parliament's Speakers Commission on Digital Democracy on Making Laws In The Digital Age.[16]

- Aspects of the research were also discussed at invited presentations to the Law Institute of Victoria in Melbourne and the Sinch Legal Technology Conference in Sydney.

- Co-authors from Cornell University presented parts of the work at a seminar in the Research School of Computer Science during a visit to Australia.

- The work was disseminated by engagement with Australia's Parliamentary Counsel's Office.

---

[15]*Is it Good Enough for the Law to Be Written for Lawyers?* https://blog.law.cornell.edu/voxpop/author/michaelcurtotti/ accessed 25 September 2015.

[16]Michael Curtotti, Submission to the UK Parliament Speaker's Commission on Digital Democracy - Making Laws in the Digital Age. 13 March 2013. http://www.parliament.uk/documents/speaker/digital-democracy/Digi007_Michael_Curtotti.pdf

# Reflections on a Body of Research

The first section of the chapter reviews each paper forming part of this research. This review also includes a review of Section 2.1 on the nature of law as that section constitutes an additional contribution of this thesis. In reviewing each paper, a brief description is provided and aspects of the paper that represent its most distinctive contribution are identified and contextualised within the overall body of research. Rather than presenting the papers in chronological order, they are presented in two broad thematic streams. Papers which deal primarily with the communication of law as text or natural language are first reviewed. Following this, papers primarily concerned with the communication of law using visualization are examined. Finally, further comment is offered on Section 2.1 of this thesis.

The second section of this chapter provides further reflections and synthesis of the work reported by this thesis as a whole. Limitations of the work are discussed and conclusions and possible future research directions are presented.

## 4.1   A Review of Published Papers and the Nature of Law

*"Of making many books there is no end, and much study wearies the body."*[1]

### 4.1.1   Corpus Based Classification of Text in Australian Contracts [Curtotti and McCreath, 2010]

*Corpus Based Classification of Text in Australian Contracts*[2] applies machine learning and hand coded rules to automatically classify text in legal contracts. More broadly the work of this paper falls within the problem of developing systems for enhancing the usability (for machines) of natural language contracts. The results of the paper were conceived as potentially contributing an input to ambiguity detection in the drafting

---

[1]Ecclesiastes 12:12, New International Version
[2]This paper is not included in this thesis, but may be accessed at http://ssrn.com/abstract=1885490

of contracts. The paper is therefore also within the domain of natural language processing.

It approaches this task by seeking to classify text within legal contracts according to the legal function of given text within a contract. Each line of text in a contract is classified by functional category. For example, which parts of the text constitute legal rules? Which parts constitute headings, execution blocks or defined terms? Three methods are applied to this task and assessed for their relative performance. The first method is application of supervised learning using machine learning techniques. The second method develops and applies hand-coded rules for the classification task. The third method looks at a hybrid classification method, using both machine learning and hand coded rules in combination.

A central problem of machine learning in the supervised case is the need for human labelled data. This data is costly to generate. However (as also established by [Curtotti and McCreath, 2010] in this context) increasing the amount of labelled data increases accuracy. A software developer therefore faces the problem of trading off accuracy against the cost of human data labelling. This cost can be reduced by hybrid methods which leverage hand coded tagging as an input for machine learning. Essentially hand coded rules can encode human knowledge about the legal domain which are readily amenable to representation by simple rule sets. For example in contract clauses the word "means" is highly associated with the presence of a legal definition.

The work on the three methods leads to a suggested methodology for functional tagging consisting of three steps:

**a.** identify the required level of accuracy;

**b.** develop hand coded rules for automatic tagging, if required accuracy is attained machine learning is unnecessary;

**c.** otherwise combine the output of the hand coded rules with other features as input to machine learning.

The results of the research demonstrated that hybrid methods were more accurate than either machine learning alone (+ 5%) or hand coding of rules alone (+ 3.7%).

In addition to its own results, work associated with this paper established technical foundations that were subsequently used in the research process. Both hand coding (for example specifically for definition detection)[Curtotti et al., 2013] and machine learning (e.g. for readability detection)[Curtotti et al., 2015b] were subsequently used in the research. This paper also suggested areas of work, including

more careful investigation of the nature of legal language.[3]

The contribution of this paper was to develop, demonstrate and test methods for functional classification of text in legal contracts. It identified a methodology, based on hand coding, enhanced where useful, by machine learning. The paper has been cited 5 times by other researchers.

### 4.1.2   A Corpus of Australian Contract Language [Curtotti and McCreath, 2011]

*A Corpus of Australian Contract Language*[4] undertakes an investigation of the nature of legal language in legal documents such as contracts. The paper applies the methods of corpus linguistics to assess and study the corpus of Australian Contract Language referred to in Section 4.1.1.

The paper undertakes description and analysis of the corpus at a number of levels. A general statistical profile is developed and compared with other English corpora to establish the differential characteristics of the legal language corpus (for example type to token ratio). The corpus is shown to conform to a Zipfian distribution (i.e. a power law distribution) in respect of word frequency. Among results was a finding that a number of characteristics of legal language follows a log normal distribution (document size, vocabulary size). Three measures (ranking by log likelihood, differences in absolute value, and frequency ratio) are assessed as tools for studying the distinctive vocabulary of contract language. Log likelihood measures and frequency ratios are compared. A simple ratio of corpus frequency showed promising results for extraction of distinctively legal terms. Log likelihood measures are particularly useful in identifying differences in functional word usage. Chunk analysis of grammatical phrases in the corpus was also undertaken. The analysis established a much higher prepositional phrase occurrence in legal sentences. Such long chains are associated with potential ambiguity and language complexity.

The contribution of this paper has been to further the characterisation of legal language within contracts. This characterisation provides insights which are valuable for natural language processing and potentially for applications such as improving the readability of legal language. Readability of legal language is the principal focus of a number of the papers discussed below. The paper has been cited 4 times by other researchers. The paper has been published in the proceedings of the 2011 International Conference on Artificial Intelligence and the Law, the leading cross-disciplinary conference addressing application of computational tools to the law. The

---

[3]Note that this paper was written while I was enrolled in a Master of Philosophy. This research provided the basis for transfer to a PhD program.

[4]Appendix A.7

### 4.1.3   A Right to Access Implies a Right to Know: An Open Online Platform for Research on the Readability of Law [Curtotti and McCreath, 2013]

*A Right to Access Implies a Right to Know: An Open Online Platform for Research on the Readability of Law*[5] furthers work towards investigating and improving the readability of legal rules. While the work reported above addresses legal rules in contracts, this and the following papers extend investigation to legal rules in legislative documents. The paper pursues readability within the broader framework of "access to law". The paper reviews established measures for readability and readability assessment in relation to legal language. It establishes that the state of the art suggests that current readability metrics are inadequate to the task of assessing readability of legal texts, and legislative sentences in particular. It also reviews previous research on the readability of legislation, finding that the research literature establishes that legislative texts are either very difficult or incomprehensible for most audiences. The paper also reports the development of an online platform for readability research on law. The platform is described above in Section 3.4. The platform is applied to undertake an initial investigation of a corpus of legislation. This corpus (popular Australian legislation) is compared with a corpus of graded readers and the Brown University Standard Corpus of Present-day American English. The study investigates a number of questions.

(a) Do traditional readability metrics or surface features of a sentence assist in assessing a sentence?

(b) Do parts of speech or chunk tagging assisting in assessing the readability of a sentence?

(c) Do such measures provide a measure of whether legislative English is 'normal' English?

Machine learning was used to investigate these questions. The graded corpus was used to establish that readability metrics were of limited value in distinguishing the reading difficulty of a sentence. The addition of speech or chunk tagging enhanced readability classification but still at a low accuracy level (30.4%). The three way comparison of the Brown, graded readers, and legislative corpus suggested a notable difference between the legislative corpus and the other corpora. Classification

---

[5]Appendix A.4

was used to show that machine learning far more readily distinguished legislation as opposed to other corpora. Comparative classification was also used treating legislative as an English 'genre'. The best classification results for the legislative corpus achieves an F-measure[6] of 0.85 as compared to the most distinctive Brown genre at an F-measure of 0.44 (i.e. legislative text is an 'outlier' as a genre of English). Also principal components analysis is used to extract and visualize principal components, strengthening the conclusion of difference between legislation and other English genres.

The work reported in the paper strengthens conclusions as to the difference of legislative language. Through an analysis of the research literature it establishes the conclusion that readability of legislative data is poor to incomprehensible for many audiences. It establishes a platform for carrying out readability research. The paper proposes and opens the pathway for subsequent collaboration with other researchers. That collaboration arose and is reported in subsequent papers. It presents a model for collaboration with a major legal publisher for the purpose of undertaking readability research. The paper was contributed to the first issue of the Journal of Open Access to Law, a then new publication of the Free Access to Law Movement, established in 2013 to provide a venue for formal publication of academic work on open access to law.

### 4.1.4 Citizen Science for Citizen Access to Law [Curtotti et al., 2015c]

*Citizen Science for Citizen Access to Law*[7] builds on *A Right to Access Implies a Right to Know*. It follows up the possibility which the previous paper identified of collaboration for the purpose of readability research on legislation. This collaboration emerged as a result of the presentation of the previous paper at the 2013 Law Via the Internet Conference and was undertaken with collaborators at the Cornell University Law School Legal Information Institute (LII). The collaboration unfolded over a period of over a year in a number of phases. The first phase involved project design and coding. In the second phase data collection was undertaken. In the third phase data analysis was undertaken. Finally the results were presented at the 2014 Law Via the Internet Conference. The project platform was deployed through a standalone server commissioned by Cornell LII and used for data provision and data collection. Javascript and PHP scripts embedded data collection within legislative pages

---

[6]The F-measure metric is widely used in machine learning and essentially represents a balance between the proportion of positive identifications which are correct ("precision") and the proportion of positive identifications actually made from all those that could have been identified ("recall"). Precision is $\frac{TP}{TP+FP}$, Recall is $\frac{TP}{TP+FN}$, and F-Measure is $\frac{2PR}{P+R}$, where $TP$ is true positives, $FP$ is false positives, $FN$ is false negatives, $P$ is precision, and $R$ is recall.[Curtotti and McCreath, 2010]

[7]Appendix A.2

at LII.[8] Figure 4.1 provides an overview of the project design. The project involved collaboration between computer science researchers, web publishers and technical staff.
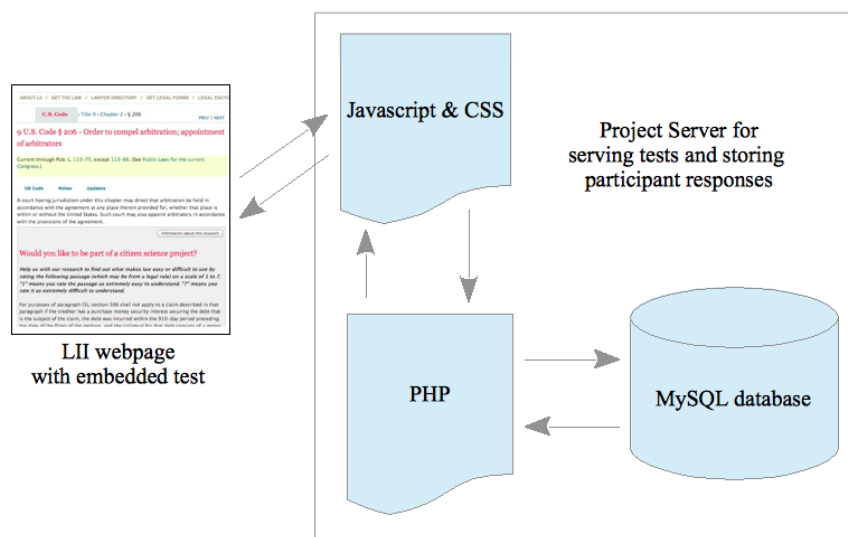


Figure 4.1: Citizen Science - Project Design

The design phase implicitly and explicitly raised a range of questions given the limited availability of research tools used in previous efforts applying crowdsourced methods to assess readability.[9] As a citizen science project, the design was framed as a largely top down project in which the "scientists" engaged "citizens" to assist with data classification. Given the end goal of improving readability of legislative materials - the unit of analysis chosen was the sentence. Approximately 1000 legislative sentences and 400 non-legal sentences were included in the study.

The study pursued three different methods for leveraging crowdsourced readability assessments of these test sentences: Likert tests, semantic differential tests and cloze tests. Each test form presented advantages and disadvantages and allowed a process of learning which can inform future research. Likert tests (requiring a user to express level of agreement with a statement concerning readability of a sentence) and semantic differential tests (measuring a user's response to adjective pairs - such as readable/unreadable) both provided subjective measures of readability. Cloze tests (based on the ability of a user to guess a missing word in a sentence) provided an objective measure of readability.[Greene, 2001; Wagner, 1986] Likert tests induced the

---

[8]Cornell LII Website https://www.law.cornell.edu/.

[9]No previous research applying crowdsourced methods to readability of legislative materials was identified.

highest response rate in an online environment, but also suffered from demographic group subjective effects (such as user desire to agree with tester which varied by demographic group). Due to inter-rater variance, in order to develop an accurate measure of the readability of a test sentence, multiple Likert tests (in the order of 20 such tests) were required. Semantic differential tests enabled the testing of different concepts related to readability of a sentence: e.g. usability and complexity. Also some of the semantic pairs enabled an assessment of whether an individual's agreement with the content of a law affected the user perceptions of its readability. The results from semantic differential testing point to a strong overlap in concepts such as clarity, readability, usability, helpfulness, familiarity and simplicity. Interestingly, the concept of clarity (measured on an 'obscure-clear' scale), appears to be most representative of this collection of characteristics, raising questions as to 'what' should be measured when seeking to improve legislative expression (i.e. should "clarity" rather than "readability" be measured in this kind of research). The semantic differential tests however suffered from a low response rate. Cloze tests although providing an objective measure, were affected by sentence length, an issue which resulted in significant difficulties in undertaking subsequent analysis. As well as comparative analysis of these testing regimes, results were combined using principal components analysis to produce a final readability measure for each test sentence. This final measure was used as an input feature for machine learning and to carry out initial machine learning, a line of investigation which was continued in a subsequent paper.

In addition to the study of readability the paper also studied the users of legislation in an online environment. (The first time such a large scale study has been published as far as we are aware). Users were invited to submit demographic data, enabling a picture to be constructed of users of the LII legislative site. The majority of respondents were non-legal (providing evidence that in the 21st century a substantial proportion of users of legislation online are not legally trained). Other findings of the demographic study included that legislation is easier for legal than non-legal audiences, and that women, individuals without tertiary education and Spanish speakers were under-represented in the context of the U.S. population. A study of Google Analytics data for the legislative pages at Cornell LII over a 12 month period also established that the readership of individual sections of legislation follows a power-law distribution. This insight is previously unreported, as far as known to the authors, and enables identification of candidates for readability improvement because of their high traffic, or alternatively for legislative pruning, because of their limited practical use. Further the research established the viability of the long term collection of readability assessments from users of legislative sites.

This paper was published in the third volume of the Journal of Open Access to

Law in 2015.

### 4.1.5    Machine Learning for Readability of Legislative Sentences [Curtotti et al., 2015b]

*Machine Learning for Readability of Legislative Sentences*[10] further extends the preliminary machine learning results reported in [Curtotti et al., 2015c]. It applies machine learning to investigate methods for enhancing the measurement of the readability of legislative sentences. As discussed in [Curtotti and McCreath, 2013] and [Curtotti et al., 2015c], existing readability metrics are inadequate for measurement of the readability of legislative text and particularly legislative *sentences*. [Curtotti et al., 2015c] developed measures of the readability of legal sentences by inviting user ratings of such sentences. This approach is extremely resource intensive and time consuming, and ideally would be replaced with the automatic measurement of readability through techniques such as machine learning. [Curtotti et al., 2015b] investigates whether particular language features improve the accuracy of machine learning for predicting the readability of a legal sentence. The paper also reports a correlation study of natural language features and reading difficulty of legislative and non-legislative sentences. The results form the basis for framing practical recommendations for improving the readability of legislative sentences.

The paper specifically investigates whether the addition of either letter ngrams or word ngrams and dependency grammar characteristics improves machine learning accuracy. Highest machine learning accuracy attained was 76.7% (73.8% for legal sentences) based on a combined set of features including word ngrams, grammatical phrase counts, features derived from a dependency grammar of the sentences. Essentially, for readability assessments, it is better to use more features, rather than less. This implies that it is indeed valuable to carry out full parsing of sentences. The paper provides insights on the distribution of the readability of sentences (essentially approaching a normal distribution), with most data close to the 'average' readability.

The correlation study reported in the paper identifies a number of natural language features that are associated with language difficulty. It identifies those features that are more characteristic of language difficulty for legislative sentences, as compared to non-legal sentences. For example prepositional phrase depth and type to token ratios are better indicators of legislative language difficulty, while sentence length is a lower ranked predictor of legislative language difficulty, particularly as compared to non-legal sentences. Also, readability metrics are confirmed to be generally worse predictors of reading difficulty in the legislative case than other charac-

---

[10]Appendix A1

teristics. The correlation study also establishes that the prediction of the readability of legislative sentences is a more complex or difficult task than predicting the readability of non-legal sentences. (i.e. correlation for any single predictor is close to 30% lower)

The findings of the paper were translated into a number of recommendations for legislative writing including avoiding high numbers of prepositions (6+); keeping sentence length below 30; keeping cross referencing to a minimum and using lexical diversity. The foregoing results are new contributions to the research literature.

This paper was published in the Proceedings of the 2015 International Conference on Artificial Intelligence and the Law and presented at the corresponding conference.

### 4.1.6   Enhancing the Visualization of Law [Curtotti and McCreath, 2012]

*Enhancing the Visualization of Law*[11] begins the second primary thread of investigation reported in this thesis. Whilst the papers above are concerned with text and its analysis, this and the following three papers described below, are primarily directed to the visual communication of law as opposed to the textual communication of law.

How, in practice, is law visualized (particularly in online environments)? This is the central question of this paper. As far as known to the authors, this paper is the first to undertake such a survey. Of course, the visualization of law is primarily concerned with presenting texts to readers. Text is a symbolic representation of spoken utterances. At a simplistic level, text is processed sequentially as a linear input of tokens and is a symbolic metaphor for speech, which is communicated sequentially through hearing. However, the presentation of text occurs within a visual context. Visual perception is both parallel (human vision instantaneously perceives a 3 dimensional world) and linear (the dimension of time). Although text continues to dominate the communication of law, visual elements have long been used, and have continued to evolve, as law has shifted from physical documents to online websites.

Examples of the use of visualization to improve legal communication are provided by Victorian reforms of the law which introduced features such as headings, sub-paragraphing and numbering.[Bowers, 1980] White space and line breaks are not meaningless in legal documents. Visualization has continued to be a concern of reformers down to the modern day. The arrival of law online can be described as having been attended by an evolutionary diversification of the visualization of law. A variety of well established legislative sites are in operation. Others are experimental. Some represent the investment of massive institutional resources. Others result from the availability of data and effective hacking by an individual or a small team.

---

[11]Appendix A.6

The simplest online visualizations at the time the paper was written, consisted of little more than an upload of text. Some visualizations were firmly rooted in a "law as document" paradigm: with online versions representing the traditional print copy of a statute. Visualizations thus focussed on visual icons of "print" versions according to timed "compilations" of the statute concerned. Other sites conceptualised law as networks, each section a node with its own page and provided associated visualization tools for navigating (the edges) between sections (nodes), to definitions, cross-references and tables of contents. Some sites conceived of law as data - first creating an enriched data representation of the law - marked up to enrich the textual information. This enriched data provides a basis for enhancing visualizations. Geographic and timeline information is incorporated into some visualizations, together with footnoting and other tools to assist users. Some sites focussed on the use of colour, fonts and layout to improve the presentation and usability of legislative information. Experimental and specific purpose visualizations strike out in different directions. For example, one visualization of bills before the U.S. Congress provided colour coded topic icons to assist in the thematic navigation of new bills. Another example was word clouds to facilitate an understanding of the actual content of a law, as opposed to the publicly stated purpose of the law. Word clouds are also used by some sites to summarise legislative content. More rarely, visualizations actually use images to represent the *meaning* of legal rules. A prominent example is drawn from transport legislation which provides images of traffic signs and to illustrate the lawful use of roundabouts.

The paper concludes with a discussion of theories of information visualization and knowledge visualization. Drawing on work of scholars in these fields the paper suggests a model for information visualization. In summary, applying prior work in the existing literature this model proposes the assessment of a visualization of legislation 'V' on the basis of a tuple of characteristics including the legislation, the user, the tasks the user is seeking to complete, a set of visualization features and comparison with a set of existing visualizations $V_2$ to $V_n$. The key contribution of this paper is the systematic review, collation and reporting of the state of the art in legislative visualization. The paper has been cited 4 times by other researchers. It was presented at the Twentieth Anniversary Conference of the Free Access to Law Movement. It thus reached the scholarly community most concerned with the presentation of law online. It has been downloaded 258 times on SSRN and viewed 39 times on Academia.

### 4.1.7   Software tools for the visualization of definition networks in legal contracts [Curtotti et al., 2013]

*Software Tools for the Visualization of Definition Networks in Legal Contracts*[12] reports practical work in the field of visualization. The work in [Curtotti and McCreath, 2010] and [Curtotti and McCreath, 2011] are a direct precursor for this work. The natural language processing tools and capacities developed in that work is applied to process natural language as input to the various visualizations which *Software Tools ...* explores. (See Figure 3.2 which illustrates the visualization pipeline.)

The paper particularly develops visualizations of the relationships between definitions in legal contracts by developing four prototype visualizations to enhance the communication and usability of such definitions. As contracts grow in length and complexity a significant proportion of the contractual language is devoted to crafting definitions supporting the legal rules in the text. These definitions can easily become problematic in themselves, or can be used strategically in order for a drafter to gain advantage for their client. Enhancing the visualization of definitions enables users to reduce errors and to prevent unintended consequences arising from definition usage. Definitions are represented in networks or mathematical graphs in which each node is a definition. These nodes are embedded in a larger set of nodes representing each section of the contract. Edges between nodes represent the presence or strength of presence of a definition within another definition or within a section. This results in a directed graph.

The first visualization produced as a result of this work, however, seeks to enhance word clouds for the representation of the content of legal contracts. Definitions are used to produce a "Definition Cloud". Definition size represents the strength of usage of a definition in a contract, while colour (using a traffic light visual coding) is used to represent hidden definitional content. A definition which links to many subsidiary definitions is represented as red because it contains hidden meaning which may mislead a reader as to the drafter's intent. A second visualization combines edge data and textual content to provide each definition with a pie chart icon which indicates how much of the real text of the definition is hidden and how important the definition is to the contract by showing numeric usage in the contract. A further visualization is focussed on the dynamics of contract reading - providing users with navigation tools allowing them to follow multi-layered definitional relationships. A more traditional node and link diagram shows the inter-relationships between defined terms used in a diagram. A final visualization uses a bimodal heat map matrix representation to visualise the strength of relationship between particular sections

---

[12]Appendix A.5

and particular definitions. All of these visualizations are novel in respect of definitions in contracts, as is a tool that is able to take natural language contract text and produce the visualizations. (See Figures 2.7, 3.2 and 4.2 for illustrations of these visualizations).

> 6.2 The **Parties** hereby grant each other a royalty free, non-exclusive licence to use each others **Existing Material** provided for the
> Existing Material means any **Material** that exists at the commencement of this **Agreement** and is used by the **Scholarship** Student for the
> Material means all material in which **Intellectual Property Rights** can consist. It includes but is not limited to, text, documents,
> Intellectual Property Rights includes all copyright and neighbouring rights, and all rights in relation to inventions (including patents), plant
> varieties, all rights in relation to registered and unregistered trade marks, registered designs and confidential information (including trade
> 6.3 secrets and know how) and circuit layouts and all other rights resulting from intellectual activity in industrial, scientific, literary or artistic
> fields.

Figure 4.2: Dynamic visualization tool for multi-layer navigation of definitions

The paper has been downloaded 112 times on SSRN and viewed 171 times on Academia. The paper was presented and published as a research abstract at the 2013 International Conference on Artificial Intelligence and the Law.

### 4.1.8 Making the Meaning of Contracts Visible - Automatic Contract Visualization [Passera et al., 2014]

*Making the Meaning of Contracts Visible Automatic Contract Visualization*[13] deals with the problem of automating the visualization of contract provisions. It builds on previous work by Passera and Haapio in the contract visualization outside the computational field [Haapio, 2011; Passera, 2012; Haapio and Passera, 2012; Haapio, 2013] and my contribution to it builds on my work in [Curtotti and McCreath, 2012] and [Curtotti et al., 2013].

The paper reports and reflects on work towards a prototype automation tool for visualizing selected contractual clauses. It notes the potential benefits of visualization in an era of information saturation. Further, automating visualization of contract terms potentially would make the creation of such contract visualization far more accessible for contract drafters - who typically lack the necessary design skills.

The paper is confined to examination of common clauses in business-to-business contracts. Three visualizations are prototyped: visualization of contract term; visualization of calculation of payment terms; and visualization of the application of liquidated damages in a contract.

The results of the work reported in the paper demonstrate the feasibility of automating contract visualization and shows how approaches already widely adopted

---

[13]Unlike other papers discussed in this chapter, I was not the principal author of this paper. My primary contribution was through sole responsibility for software development, participation in researcher collaborative meetings and paper editing. The paper received a LexisNexis Top 10 Prize at IRIS 2014.

in the legal industry in regard of document assembly and data capture can be readily extended to automating visualization. In contrast to [Curtotti and McCreath, 2012], natural language processing is only used to a minimal degree in the application, leveraging user input instead as a means of constructing visualizations and using text libraries to compile contract clauses, given user input. An insight from the paper was the identification of the current absence of a visual language adequate for representing some types of contract rules. Many rules do not lend themselves to ready visual representation. The development of a visual language (as exemplified with creative commons visualization of legal rules) is necessary to extend visualization to a broader range of legal provisions.

Among the insights reported in the paper are insights drawn from iterative creation of such clause visualizations. The process of clause visualizations provided a feedback mechanism which iteratively improved the textual drafting to minimise ambiguity.

The contribution of this paper was recognised in being awarded a Lexis Nexis Top 10 Papers Prize at the conference at which it was presented. The paper has been viewed 343 times at Academia and downloaded 36 times at SSRN.

### 4.1.9   Interdisciplinary Cooperation in Legal Design and Communication [Curtotti et al., 2015a]

*Interdisciplinary Cooperation in Legal Design and Communication*[14] extended the reflective work initially explored in [Passera et al., 2014]. This paper reflected on the implications of cross-disciplinary engagement with law, drawing on the practical work described in [Passera et al., 2014]. A much broader readership for law has emerged in the 21st century as law has been made available online. Computer scientists, web designers and others have necessarily engaged with the process of communication of law in this context. The ways in which different professions think of the law was highlighted by the experience of the researchers. Designers approach law as designed artefact and, typically, with a view to empowering and enabling the end user of the artefact. The functionality, usability and user experience are all relevant from this viewpoint. Computer scientists and software engineers bring their own paradigms - focussing on law as data - but also adapting and applying disciplinary paradigms their engagement with law. Thinking of law as part of a software product is one example. Application of principles of abstraction is another. Business managers engaging with contracts are more concerned with them as enabling instruments as much as with risk mitigation. As an exercise in cross-disciplinary cooperation, the work was

---

[14]Appendix A.3

in the field of knowledge visualization rather than information visualization. How lawyers are educated is strikingly different to the education of software engineers - the latter being immersed in their stock in trade "computer code", whereas law students traditionally will rarely, if ever, see a contract in their core legal education. The nature of law itself becomes open to new questions from this multi-disciplinary viewpoint. A design paradigm carries implication for traditional models of law - and even for the core characterisation of laws as *constraints* on human behaviour. Rather laws can be seen as enabling, rather than constraining. The legal profession itself is open to reinvention, when such paradigms are taken into account. This paper was presented in Europe at the 18th International Legal Informatics Symposium IRIS 2015.

### 4.1.10 The Nature of Law

This section, in contrast to the above, does not report a published paper, rather referring to Section 2.1 of this thesis. It is however relevant to do so here, as Section 2.1 essentially represents a final extension of the investigations carried out as part of this thesis. The outline of questions raised in [Curtotti et al., 2015a] as to the nature of law is explored in greater detail and within an overall theoretical framework for that investigation. As the issues have already been covered in some detail above, there is no need for an extensive reiteration at this point. The additional contribution of this section of the thesis is as follows. The nature of law is investigated from a number of additional perspectives, including as network and complex adaptive system. Each investigated perspective offers conclusions as to the nature of law. This enables a coherent, multfacetted description of law, avoiding oversimplification. The "law as ..." framework is adopted but also further clarified as a manner of investigating law - re-conceptualising it as a means for carrying out different thought experiments for investigating the nature of law. In particular the conclusion is offered that the framework enables law to be investigated from numerous perspectives, and that these perspectives can be combined to offer an enriched understanding of the phenomenon of law - rather than competing and mutually exclusive versions of "legal truth". As the investigation is multidisciplinary in nature it queries the influence of now dated scientific paradigms which underpin how law has been traditionally understood.

## 4.2   Reflections and Synthesis: Enhancing the Communication of Law

*"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."*

Bilbo Baggins, J.R.R. Tolkien, The Lord of the Rings

### 4.2.1   Reflections and Synthesis

The foregoing has highlighted the individual contributions of each paper. Taken as a whole the work reported in this thesis reflects a multifacetted contribution to enhancing the communication of law with particular focus on application or use of computational tools. The scope of work has covered both legislative "rules" and contract "rules".

Basic work characterising such rules as linguistic data is advanced through the application of corpus linguistics to legislative texts. Natural language processing, machine learning and hybrid application of hand coded rules are applied to advance functional classification of text within contracts. The textually focussed investigation substantively addressed readability. Existing work in the field is collated and described and used as a foundation for advancing the state of the art. As far as readability of law is concerned, principles advanced by the plain language movement have become widely accepted practice in the legal field. However, a review of existing research established that the central goal of plain language in application to legislation, has not been achieved. Legislative language continues to remain largely inaccessible to large audiences which seek to use it - particularly in an online context. Work in this direction develops quantitative indicators of the demographic characteristics of the audience that seeks to read law online (i.e. the audience for legislative communication in a digital age). This is the first large scale work of this kind, as far as I am aware. In addition, the research undertakes the first large scale use of crowdsourced citizen science to assess the readability of legislative material. The collection of thousands of such assessments is unprecedented. The assessments were used, after application of machine learning, to develop indicators for improving legislative readability. The readability research also investigated the development of methods for readability assessments using online crowdsourced assessments. Three measures were investigated: Likert tests; semantic differential tests and cloze tests - highlighting the advantages and disadvantages of each. The research established the feasibility of long term data collection for readability purposes.

Work exploring the visual communication of law applied computational tools to investigate the visualization of definition networks in contracts and methods and approaches for automating the visualization of selected contract clauses. This work established proof of concept for both visualization sets. The work also included a review of the state of the art in visualization of law in online environments. It established both an environment of diverse experimentation and also, in many cases, of an inertia inherent in existing concepts of law. Law as "document" has transcended its physical manifestation in paper based documents, yet remained conceptually imprisoned in digital analogues of paper precedents. Other sites have broken away from these conceptual moorings and re-presented laws beginning from a concept of law as "data".

In addition to the theoretical results, the research has contributed to the body of knowledge in other ways. The production of the Australian Contract Corpus, and the Corpus of American Legislative English, are examples of data now available for investigation by other researchers. The Readability Research Platform represents a tool made available to researchers for undertaking readability research.

This thesis is undertaken by compilation and is multidisciplinary in character. It has explored the problem domain of "enhancing communication of law" and has contributed to knowledge concerning that domain. As research in this domain unfolded initial concepts and goals, embodied in earlier papers, were revised and direction shifted. Initially, the focus was very much in the domain of applied computational science - adapting existing tools for enhancing the drafting contracts. As the research unfolded however, it became clear that enhancing the drafting of contracts sat within a broader set of issues concerned with enhancing the communication of legal rules as a whole. Further it become evident that a better understanding of the data itself (legal language), was required if this goal was to be pursued more effectively. Early in the research process, "text" (the written word) was implicitly set aside as an insufficient embodiment of the information found in legal rules. Initially this manifested in work on visualizing definition networks in legal contracts and later the visualization of law became an explicit focus of research in the context of the online visualization of law. This in turn enabled the identification of the documentary paradigm of law as a limiting factor impeding visualization and re-use of law in novel ways. It also became evident that the questions addressed by the research required methodological foundations to be laid.

Theories of legal visualization are rudimentary - and within the computational context almost non-existent. Theories, approaches and methods from related fields needed to be extrapolated in application to legal rules. Further, the centrality of written communication - the data of legal rules - impelled an examination of how

the law was written. How is the writing of law and the enhancement of that writing to be undertaken? Tools such as existing readability metrics were explored and found to be inadequate. Such investigations established the need for tools for measuring the readability of law. But without "gold standard" data, evaluating any approach to enhancing readability would not be progressed. This led to a consideration of the need for crowdsourced user studies and the application of citizen science in this new field. As the research unfolded it opened opportunities for collaboration. Initial collaboration within the ANU expanded to cooperation with researchers in the United States (in respect of readability) and in the Europe (in respect of visualization). And by its nature, collaboration provided a rich environment for further learning.

In the final stage of the research the nature of law itself came into focus. Innocuous and familiar to the researcher, through many years of legal experience - assumptions as to its character became open to question. The legal "rule" perhaps was something more than it had meant to the lawyer - or the conceptual model that law school education had prioritised: law as "command" and its close derivates. What, indeed, was being enhanced, when enhancing the communication of law is spoken about? If law is not merely a "command" but - as it was carefully investigated - a chameleon-like entity perhaps better thought of as a dynamic adaptive system - the enhancement of communication conceived as "clearer" commands was completely inadequate. In this inadequacy is the potential for enhancing not just the communication of law, rather the opening of new avenues of researching the law itself. The "law as" paradigm which is applied in Section 2.1 suggests a framework for investigating law from a multidisciplinary perspective. The paradigm is interpreted not as telling us what the law "is" - but rather as an experimental metaphor that allows us to investigate the law in many different ways. Each investigation adds to our understanding of the phenomenon of law - a complex phenomena which is not adequately captured by theories which seek to simplify - and reductively define it - in one frame. As discussed, seeing the law in these new ways - "as" (for example) "network", enables us to envisage new ways in which the law can be accessed, used and made more useful and usable to those who engage with it.

The way-markers of this research journey - the specific research outcomes of each paper - are contributions in their own right. Taken together they contribute to the body of knowledge concerning the communication of law, its enhancement, and the character of the matter to which improvement is sought. These way markers have been outlined above. Prominent in the research outcomes has been an application of computational tools and methods, a starting point and constant of the research. These computational tools and methods are outlined in Chapter 3.

The "law" is something that is known in virtually every human society. Since the

agricultural revolution, law has been closely associated with text and documents - whether that document was created using stone, baked clay, rice piper, papyrus or wood pulp. We stand at a threshold of a revolution for the law. It has broken free of "document" - though that fact is as yet little known. Of course, computers and the information revolution have made this possible. As we have seen this revolution opens the very character of law to new investigation. It has changed who uses the law. It has made enhancing the communication of law both possible and essential to those who now use it. Indeed, this implies another revolution in the law of which we are still barely conscious: the language of the law has escaped the exclusive control of the legally literate. It is in the nature of language that its reading audience will over time powerfully influence how it is expressed.

The research reported here does not seek to present final conclusions as to how the communication of law is to be enhanced. It reports the results of research that has explored this question in a number of dimensions. It is highly likely the information revolution will continue to drive improvements in the communication of law - whether in directions envisaged here or in other forms. If that enhancement is to be systematised it requires the development of a coherent body of knowledge that addresses that enhancement. The research reported here undertakes such a systematisation, in the context of the application of computational tools to enhancing communication; and in the context of the specific investigation of the communication (including visualization) of legislative and contractual language.

### 4.2.2　Limitations and Future Research Directions

There are a number of limitations to this work. These are stated below.

**Language:** The research reported here, as far as concerns textual communication, is limited to English. The conclusions reached as to readability are not necessarily extensible to other languages, although it is interesting to note that research in relation to Italian legislation reaches similar findings in respect of high prepositional use. (See [Venturi, 2008]) A useful direction of future research would be either the comparison of research reported here, or its extension, to other language contexts. In addition to considering how the language context may make a difference to measuring readability - it may be feasible to identify meta-language characteristics that remain invariable or true irrespective of the language in which a law is written. In other words, is there something specific to the nature of the legal domain, that transcends human language and affects readability irrespective of language.

**Jurisdiction:** A further limitation of the research is its focus primarily within the Australian jurisdiction (in respect of contracts) and the United States (in respect of

legislation). The writing of legal materials varies by jurisdiction. For example U.S. contract and legislative style is noticeably different to Australian contract and legislative style. This is true in respect of two comparatively close jurisdictions. Differences may be even greater for other jurisdictions, and represents another limitation of the research. A potential direction for future research is extension of readability studies to a greater diversity of jurisdictions - including potentially sub-national units in respect of federal or quasi-federal states.

**Gold Standard Readability Data:** Machine learning for classification of data depends on the existence of a "gold standard" - generally a human tagged dataset as training data for classifying previously unseen instances. The research included the application of crowdsourcing to undertake user driven assessments of readability of legislative sentences. The research collected tens of thousands of such assessments. Even at this scale - the number of tagged instances which resulted allow only limited accuracy for machine learning purposes. This is due to the power law distribution of linguistic features. However, methods for overcoming this limitation are identified as a result of this research - i.e. the feasibility of carrying out long term (multi-year) readability data collection is demonstrated.

**Citizen Science:** A component of the research reported here falls within the domain of citizen science. While the research is an instance of a large scale application of citizen science to the problem of enhancing the communication of law, the citizen engagement was largely in the form of data assessment and data collection. Users assessed sentences for readability and provided data about themselves. This engagement of citizen scientists is only one of a number of possible levels over which citizen science may range. For example, citizen scientists may be engaged in the process of research formulation, at one end, up to and including involvement as co-authors of resulting research. Future research may explore a more extensive engagement of citizens in enhancing the communication of law which affects their lives as citizens.

**Non-Legislative Materials**: The results reported here are inapplicable to other legal genres (for example judicial decisions). Such materials require their own investigation and measures of readability. As far as the author is aware, computational techniques have not been applied to investigate the readability of other legal genres.

**Investigating Re-imagined Law:** An insight drawn from the research was the need to query the nature of law. As discussed above, the research leads to re-conceptualisation of law using the 'law as' paradigm and a methodology for the investigation of law as a complex social phenomenon. These results however came at the tail end of the research. Accordingly, a limitation of the research, is that much of it proceeds with the implicit assumption that laws are rules, or at most document, language or data. Future research, applying computational techniques, could inves-

tigate law in different ways drawing on its broader conceptualisation. For example - insights drawn from examining law as a complex adaptive system - could investigate and apply computational tools to the dynamics of legislative evolution in domains such as tax or migration policy. Computational tools, combined with the public availability of point in time data, now make such investigations more feasible than would have been true in the past.

**Investigating Law as Empowerment Rather than Constraint**: An insight noted in Section 2.1 is that law has been traditionally primarily conceived as a set of constraints on society. Both feminist and design perspectives call this conception into question. Power, for example, is not solely a phenomenon that constrains. Law, need not be thought of through this lens either. What law as empowerment might mean, is a thought experiment that represents another potential line of investigation.

**Investigating the Communication of Knowledge**: The readability research reported here, as noted in Section 2.1.5, only assesses the readability of law as "information". That section noted the *data –> information –> knowledge* model associated with communication. The readability research reported here addresses whether information is conveyed to a reader (subjectively and objectively). It does not address whether real "knowledge" of the law is achieved in the reader. i.e. whether the users' subjective belief that they have "understood" a legal sentence - equates to real knowledge of its communicator's intent. This is a field of potential further investigation, particularly developing computational tools for such investigations.

**A Visual Language for Legal Communication**: One of the limits encountered in seeking to visualise legal rules, is the absence of a visual language for such communication. This suggests the need for the generation of a common visual language for legal communication, if the full potential of visualization is to be realised. The investigation of the generation and application of such a visual language is a further area of potential future investigation.

# Abstracts and Visual Summary of Publications

This chapter provides a summary of publications. Traditionally this is undertaken by presenting abstracts of each paper, which is done here. In addition, to more immediately communicate content and focus, a visual summary is also provided, using visualizations created using the Wordle website.[1]The visualizations are basic word clouds highlighting the prevalence of terms used in each paper. Colour and font are features of the visualization that contribute to the aesthetic effect of the visualization, but do not of themselves convey meaning.

---

[1]Wordle - Beautiful Word Clouds http://www.wordle.net/

**A.1.** Curtotti, M., McCreath, E., Bruce, T., Frug, S., Weibel, W. and Ceynowa, N. *Machine Learning for Readability of Legislative Sentences*. Proceedings of the Fifteenth International Conference on Artificial Intelligence and the Law. (ICAIL, 2015) Association for Computational Machinery (ACM)



Figure 5.1: Machine Learning for Readability of Legislative Sentences

**ABSTRACT:** Improving the readability of legislation is an important and unresolved problem. Recently, researchers have begun to apply legal informatics to this problem. This paper applies machine learning to predict the readability of sentences from legislation and regulations. A corpus of sentences from the United States Code and US Code of Federal Regulations was created. Each sentence was labelled for language difficulty using results from a large-scale crowdsourced study undertaken during 2014. The corpus was used as training and test data for machine learning. The corpus includes a version tagged using the Stanford parser context free grammar and a version tagged using the Stanford dependency grammar parser. The corpus is described and made available to interested researchers. We investigated whether extending natural language features available as input to machine learning improves the accuracy of prediction. Among features evaluated are those from the context free and dependency grammars. Letter and word ngrams were also studied. We found the addition of such features improves accuracy of prediction on legal language. We also undertake a correlation study of natural language features and language difficulty drawing insights as to the characteristics that may make legal language more difficult. These insights, and those from machine learning, enable us to describe a system for reducing legal language difficulty and to identify a number of suggested heuristics for improving the writing of legislation and regulations.

**A.2.** Curtotti, M., Weibel, W., McCreath, E., Ceynowa, N., Frug, S. and Bruce, T. *Citizen Science for Citizen Access to Law*. Journal of Open Access to Law, 2015, Vol 3:1.



Figure 5.2: Citizen Science for Citizen Access to Law

**ABSTRACT:** The widespread availability of legal materials online has opened the law to a new and greatly expanded readership. These new readers need the law to be readable by them when they encounter it. However, the available empirical research supports a conclusion that legislation is difficult to read if not incomprehensible to most citizens. We review approaches that have been used to measure the readability of text including readability metrics, cloze testing and application of machine learning. We report the creation and testing of an open online platform for readability research. This platform is made available to researchers interested in undertaking research on the readability of legal materials. To demonstrate the capabilities of the platform, we report its initial application to a corpus of legislation. Linguistic characteristics are extracted using the platform and then used as input features for machine learning using the Weka package. Wide differences are found between sentences in a corpus of legislation and those in a corpus of graded reading material or in the Brown corpus (a balanced corpus of English written genres). Readability metrics are found to be of little value in classifying sentences by grade reading level (noting that such metrics were not designed to be used with isolated sentences).

**Keywords:** readability, legislation, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, cloze testing

**A.3.** Curtotti, M., Haapio, H. and Passera, S. *Interdisciplinary Cooperation in Legal Design and Communication*. Kooperation - Digitale Aasgabe zum Tagangsband des 18. Internationalen Rechtsinformatik Symposions IRIS 2015 (Cooperation - Digital Proceedings of 18th International Legal Informatics Symposium IRIS 2015), JurisletterIT 26 February 2015.



Figure 5.3: Interdisciplinary Cooperation in Legal Design and Communication

**ABSTRACT:** The last two decades have seen law emerge online. This development has engaged computer scientists and web designers in communicating law. Recently, serious work has begun on visualizing contract clauses, generating cooperation between designers, computer scientists, business people, lawyers and others. New insights arise from such cross disciplinary collaborations. Each discipline provides theoretical insights as to how legal design and communication might be approached. More profoundly each has the potential to recast relationships - what does it mean for the 'power' of law makers to be exercised in the context of such paradigms? How do such insights enable us to reconsider the role of lawyers: the traditional custodians of legal rules? We examine these questions from a theoretical viewpoint, and reflect on our own cross-disciplinary collaboration in the creation of a proof-of-concept tool for automation of contract visualization.

**Keywords:** legal design, legal visualization, legal communication, multidisciplinary collaboration, contract visualization, cross-professional communication

**A.4** Curtotti M. and McCreath E. *A Right to Access Implies A Right to Know: An Open Online Platform for Research on the Readability of Law*. Journal of Open Access to Law, 2013. Vol 1:1



Figure 5.4: A Right to Access Implies a Right to Know: An Open Online Platform for Research on the Readability of Law

**ABSTRACT:** The widespread availability of legal materials online has opened the law to a new and greatly expanded readership. These new readers need the law to be readable by them when they encounter it. However, the available empirical research supports a conclusion that legislation is difficult to read if not incomprehensible to most citizens. We review approaches that have been used to measure the readability of text including readability metrics, cloze testing and application of machine learning. We report the creation and testing of an open online platform for readability research. This platform is made available to researchers interested in undertaking research on the readability of legal materials. To demonstrate the capabilities of the platform, we report its initial application to a corpus of legislation. Linguistic characteristics are extracted using the platform and then used as input features for machine learning using the Weka package. Wide differences are found between sentences in a corpus of legislation and those in a corpus of graded reading material or in the Brown corpus (a balanced corpus of English written genres). Readability metrics are found to be of little value in classifying sentences by grade reading level (noting that such metrics were not designed to be used with isolated sentences).

**Keywords:** readability, legislation, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, cloze testing

**A.5** Curtotti, M., McCreath, E. and Sridharan, S. *Software Tools for the Visualization of Definition Networks in Legal Contracts*. (Research Abstract) Proceedings of the Fourteenth International Conference on Artificial Intelligence and the Law (ICAIL, 2013). Association for Computational Machinery (ACM).



Figure 5.5: Software Tools for the Visualization of Definition Networks

**ABSTRACT:** This paper describes the development of prototype software-based tools for visualizing definitions within legal contracts. The tools demonstrate visualization techniques for enhancing the readability and comprehension of definitions and their associated characteristics. This contributes to more accurate and efficient drafting or reading of contracts through the exploration of the meaning and use of definitions including via word clouds, multilayer navigation, adjacency matrix and graph tree representations.

**Keywords:** definitions, legal contracts, word clouds, network visualization, contract visualization, text visualization, graph metrics

**A.6** Curtotti, M. and McCreath, E. *Enhancing the Visualization of Law*. Presented October 2012 20th Anniversary Conference of Law via the Internet - Cornell University



Figure 5.6: Enhancing the Visualization of Law

**INTRODUCTION:** The accessibility of law has undergone a revolution in the last two decades as public good, official and commercial initiatives have made legislation (and other legal materials) accessible online. In respect of legislation, this development has followed centuries of refinement in how the law is written and presented. The presentation of the law (or in its 21st century manifestation - its visualization) has long been known to influence its readability (itself a dimension of the accessibility of law). Online legislation sites vary widely in their approaches. The most basic present legislation as a scrollable text (in practice reverting to the equivalent of a single scroll of paper), the most novel use features such as colour, graphs, images, moving pictures and information enhancement to improve visualizations. Some sites focus on providing laws as downloadable documents in various formats - emphasising the online availability of 'the official version'. Many sites provide access or links to accompanying materials such as explanatory memoranda, subordinate legislation or court interpretations. Some sites offer legal rules within legislation as navigable nodes, providing links to key information including (in some cases) links to cross references and defined terms used in a legal rule. Search tools are a basic feature offered by most sites. Some sites provide solutions which enhance visualization using selection of font, font size, content and colour. A small number of sites provide point-in-time access to legislatoin. Some research sites or approaches explore the presentation of legislation or bills in radically different forms: such as graph visualizations

or as topic colour-coded icons. In this paper we first review examples of such visualizations and highlight various approaches that are available in official and public good sites in selected jurisdictions. We then briefly present our own visualizations that focus on the enhancement of the visualization of definitions in the parallel domain of legal contracts. Such visualizations are readily transferable to the legislative domain. The welter of approaches available raises the question of how we may evaluate the utility of a particular visualization. On what basis are we able to suggest, for example, that a basic presentation of text is any worse than a site which provides graphical sliders allowing access to point-in-time versions of legislation? While we intuitively expect more 'advanced' visualizations to be preferable, what are our theoretical or empirical grounds for such conclusions? Furthermore we might ask, better or worse for whom? Current visualizations do not necessarily distinguish between lawyers, citizens, law makers, advocates and other users, who have quite distinct needs. Drawing particularly on the fields of information visualization and knowledge visualization, we conclude by presenting a potential theoretical framework for grounding the visualization of legislation, and discuss the evaluation of legislative visualization. While a variety of definitions exist of what might be meant by visualization, in the context of this paper we primarily mean the use of graphics, images or symbols (other than words themselves) to enhance the communication of meaning contained in or associated with (legislative) text. While primarily in the text itself, meaning in text extends beyond the words themselves, for example information such as document structuring or relation- ships between concepts found in text. In Section 4 we explore this definitional issue further.

**Keywords:** visualization of law, publication of law online, access to law, visualization of information, evaluation of visualization

———————————————

**A.7** Curtotti, M. and McCreath, E. *A Corpus of Australian Contract Language*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law (ICAIL, 2011). Association for Computational Machinery (ACM).



Figure 5.7: A Corpus of Australian Contract Language

**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.

**Keywords:** corpus linguistics, legal contracts, natural language processing, characterisation of legal language, contract corpus.

# Bibliography

AGUILÓ-REGLA, J., 2005. Introduction: legal informatics and the conceptions of the law. In *Law and the Semantic Web*, 18–24. Springer. (cited on pages 12 and 13)

AHDIEH, R. B., 2006. Dialectical regulation. *Connecticut Law Review*, 38 (2006), 863. (cited on page 57)

ALLEN, L. E. AND ENGHOLM, C. R., 1978. Normalized legal drafting and the query method. *Journal of Legal Education*, 29, 4 (1978), 380–412. (cited on page 24)

ANDERLINI, L. AND FELLI, L., 2004. Bounded rationality and incomplete contracts. *Research in Economics*, 58, 1 (2004), 3–30. (cited on page 44)

ANDERSON, E.; SCHEIDEGGER, C.; AND SILVA, C., 2010. The Role of Perception and Cognition Costs in Models of Visualization Effectiveness. *VizWeek 2010: The Role of Theory in Information Visualization*, (2010). (cited on page 76)

ATTENBOROUGH, F., 1922. *The Laws of the Earliest English Kings*. Cambridge University Press. (cited on page 24)

AXELROD, R., 1986. An evolutionary approach to norms. *American political science review*, 80, 4 (1986), 1095–1111. (cited on pages 54 and 55)

AXELROD, R. AND HAMILTON, W. D., 1981. The evolution of cooperation. *Science*, 211, 4489 (1981), 1390–1396. (cited on page 55)

BACK, R. J. AND VON WRIGHT, J., 2000. Contracts as mathematical entities in programming logic. Technical report, Turku Centre for Computer Science. (cited on page 45)

BARNETT, H., 1998. *Introduction to feminist jurisprudence*. Cavendish Publishing Limited. (cited on page 17)

BAYS, C., 1987. Candidates for the game of life in three dimensions. *Complex Systems*, 1, 3 (1987), 373–400. (cited on page 53)

BECKNER, C.; BLYTHE, R.; BYBEE, J.; CHRISTIANSEN, M. H.; CROFT, W.; ELLIS, N. C.; HOLLAND, J.; KE, J.; LARSEN-FREEMAN, D.; AND SCHOENEMANN, T., 2009. Language

is a complex adaptive system: Position paper. *Language Learning*, 59, S1 (2009), 1–26. (cited on page 63)

Bench-Capon, T.; Araszkiewicz, M.; Ashley, K.; Atkinson, K.; Bex, F.; Borges, F.; Bourcier, D.; Bourgine, P.; Conrad, J. G.; Francesconi, E.; et al., 2012. A history of AI and Law in 50 papers: 25 years of the International Conference on AI and Law. *Artificial Intelligence and Law*, 20, 3 (2012), 215–319. (cited on page 72)

Bennion, F., 1983. *Statute law*. Oyez. (cited on page 28)

Benson, R. W., 1984. End of legalese: The game is over. *NYU Rev. L. & Soc. Change*, 13 (1984), 519. (cited on page 5)

Bentham, J., 1843. Nomography, or the art of inditing laws. *The Works of Jeremy Bentham (ed. J. Bowring)*, 3 (1843), 231 et seq. (cited on pages 28 and 29)

Berners-Lee, T., 1989. Information management: A proposal. http://www.w3.org/History/1989/proposal.html. (cited on page 72)

Berners-Lee, T., 1991a. The original http as defined in 1991. http://www.w3.org/Protocols/HTTP/AsImplemented.html. (cited on page 72)

Berners-Lee, T., 1991b. Re: Sgml/html docs, x browser. http://lists.w3.org/Archives/Public/www-talk/1991NovDec/0020.html. (cited on page 72)

Berners-Lee, T. and Cailliau, R., 1990. Worldwideweb: Proposal for a hypertext project. http://www.w3.org/Proposal.html. (cited on page 72)

Berners-Lee, T. et al., 2006. The worldwideweb browser. http://www.w3.org/People/Berners-Lee/WorldWideWeb.html. (cited on page 72)

Bird, S.; Loper, E.; and Klein, E., 2009. *Natural Language Processing with Python*. O'Reilly Media Inc. (cited on pages 86, 88, and 93)

Bix, B., 1999. *Jurisprudence: Theory and context*. Sweet & Maxwell, London, 2nd edn. (cited on page 14)

Blakeley, G. and Bryson, V. (Eds.), 2007. *The impact of feminism on political concepts and debates*. Manchester University Press. (cited on page 17)

Blum, C. and Li, X., 2008. Swarm intelligence in optimization. In *Swarm Intelligence*, 43–85. Springer. (cited on page 54)

Boccara, N., 2004. *Modeling complex systems*. Springer. (cited on page 58)

BOISOT, M. AND CANALS, A., 2004. Data, information and knowledge: have we got it right? *Journal of Evolutionary Economics*, 14, 1 (2004), 43–67. (cited on page 33)

BOMMARITO, M. J. AND KATZ, D., 2010. A mathematical approach to the study of the United States Code. *Physica A: Statistical Mechanics and its Applications*, 389, 19 (2010), 4195–4200. (cited on pages 39 and 57)

BOWERS, F., 1980. Victorian reforms in legislative drafting. *Tijdschrift voor Rechtsgeschiedenis*, 48 (1980), 329. (cited on pages 29 and 103)

BRONSTEIN, I.; SEMENDYAYEV, K.; MUSIOL, G.; AND MUEHLIG, H., 2007. *Handbook of Mathematics*. Springer, fifth edn. (cited on page 38)

BRUNSCHWIG, C. R., 2014. On visual law: Visual legal communication practices and their scholarly exploration. *Zeichen und Zauber des Rechts: Festschrift für Friedrich Lachmayer, Erich Schweihofer et al.(eds.), Bern: Editions Weblaw*, (2014), 899–933. (cited on page 80)

BURKHARD, R., 2004. Learning from architects: the difference between knowledge visualization and information visualization. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, 519–524. IEEE. (cited on pages 6 and 76)

CARD, S. AND MACKINLAY, J., 1997. The structure of the information visualization design space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, 92–99. IEEE. (cited on pages 6 and 76)

CHALMERS, A. F., 1982. *What is this thing called science?* University of Queensland Press, 2nd edn. (cited on page 68)

CHALMERS, A. F., 2013. *What is this thing called science?* Hackett Publishing, 4th edn. (cited on pages 66 and 67)

CHEN, M.; EBERT, D.; HAGEN, H.; LARAMEE, R. S.; VAN LIERE, R.; MA, K.-L.; RIBARSKY, W.; SCHEUERMANN, G.; AND SILVER, D., 2009. Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, 29, 1 (2009), 12–19. (cited on pages 34 and 35)

CHEN, M. AND JAENICKE, H., 2010. An information-theoretic framework for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16, 6 (2010), 1206–1215. (cited on page 76)

CHUN, W., 2007. *Core python programming*. Prentice Hall Professional, 2nd edn. (cited on page 43)

COLLINS-THOMPSON, K., 2014. Computational assessment of text readability: A survey of past, present, and future research. *working draft*, (July 2014). (cited on pages 5 and 75)

CONNOLLY, D. AND CAILLIAU, R., 2000. A little history of the world wide web. http://www.w3.org/History.html. (cited on page 72)

COTTERRELL, R. B., 2003. *The politics of jurisprudence: A critical introduction to legal philosophy*. Lexisnexis UK, 2nd edn. (cited on pages 16, 17, and 57)

CRYSTAL, D. AND DAVY, D., 1969. *Investigating English Style*. L.: Longman. (cited on page 28)

CURTOTTI, M.; HAAPIO, H.; AND PASSERA, S., 2015a. Interdisciplinary cooperation in legal design and communication. In *Cooperation - Digital Proceedings of 18th International Legal Informatics Symposium IRIS 2015), JurisletterIT 26 February 2015*. International Legal Informatics Symposium (IRIS). (cited on pages xiii, 10, 12, 14, 19, 22, 31, 46, 49, 50, 87, 107, and 108)

CURTOTTI, M. AND MCCREATH, E., 2010. Corpus Based Classification of Text in Australian Contracts. In *Proceedings of the Australasian Language Technology Association Workshop 2010*. (cited on pages xii, 5, 12, 21, 85, 90, 95, 96, 99, and 105)

CURTOTTI, M. AND MCCREATH, E., 2011. A corpus of Australian contract language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law 2011*. (cited on pages xii, 4, 12, 21, 85, 86, 97, and 105)

CURTOTTI, M. AND MCCREATH, E., 2012. Enhancing the visualization of law. https://blog.law.cornell.edu/lvi2012/presentation/enhancing-the-visualization-of-law/. (cited on pages xii, 6, 12, 22, 30, 32, 39, 62, 75, 76, 78, 87, 91, 103, 106, and 107)

CURTOTTI, M. AND MCCREATH, E., 2013. A right to access implies a right to know: An open online platform for research on the readability of law. *Journal of Open Access to Law*, 1, 1 (2013). (cited on pages xii, 5, 12, 27, 86, 90, 93, 98, and 102)

CURTOTTI, M.; MCCREATH, E.; BRUCE, T.; FRUG, S.; WEIBEL, W.; AND CEYNOWA, N., 2015b. Machine learning for readability of legislative sentences. In *Proceedings of the 15th International Conference on Artificial Intelligence and the Law*, 53–62. The International Association for Artificial Intelligence and the Law, ACM. (cited on pages xii, 5, 12, 22, 26, 35, 63, 86, 87, 89, 90, 92, 93, 96, and 102)

CURTOTTI, M.; MCCREATH, E.; AND SRIDHARAN, S., 2013. Software tools for the visualization of definition networks in legal contracts. In *Proceedings of the Fourteenth*

*International Conference on Artificial Intelligence and Law*, 192–196. ACM. (cited on pages xiii, 10, 12, 21, 32, 35, 39, 62, 85, 89, 91, 96, 105, and 106)

CURTOTTI, M.; WEIBEL, W.; MCCREATH, E.; CEYNOWA, N.; FRUG, S.; AND BRUCE, T., 2015c. Citizen science for citizen access to law. *Journal of Open Access to Law*, 3, 1 (2015). (cited on pages xii, 2, 5, 10, 12, 22, 27, 29, 35, 40, 59, 63, 74, 75, 86, 87, 90, 92, 93, 99, and 102)

DASKALOPULU, A., 1998. Legal contract drafting at the micro-level. In *Law in the Information Society: 5th International Conference of the Institute of Legal Documentation (IDG) of the Italian National Research Council, Florence, Italy (On CD-ROM)*. (cited on page 45)

DASKALOPULU, A., 2000. Modelling legal contracts as processes. In *Database and Expert Systems Applications, 2000. Proceedings. 11th International Workshop on*, 1074–1079. IEEE. (cited on pages 44 and 45)

DASKALOPULU, A. AND SERGOT, M., 1997. The representation of legal contracts. *AI & Society*, 11, 1 (1997), 6–17. (cited on page 7)

DASKALOPULU, A.-K., 1999. *Logic-based tools for the analysis and representation of legal contracts*. Ph.D. thesis, University of London. (cited on page 45)

DE MAAT, E. AND WINKELS, R., 2008. Automatic classification of sentences in dutch laws. In *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, 207–216. IOS Press. (cited on page 7)

DE MARNEFFE, M. C. AND MANNING, C. D., 2008. Stanford typed dependencies manual. *URL http://nlp. stanford.edu/software/dependencies manual.pdf*, (2008). (cited on page 93)

DE NOOY, W.; MRVAR, A.; AND BATAGELJ, V., 2005. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press. (cited on page 38)

DEGÓRSKI, Ł.; MARCINCZUK, M.; AND PRZEPIÓRKOWSKI, A., 2008. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*. (cited on page 7)

DOHERTY, M., 2002. *Jurisprudence: The Philosophy of Law. Sourcebook*. Old Bailey Press, 2nd edn. (cited on pages 7, 14, and 57)

DuBay, W., 2004. The principles of readability. *Impact Information*, (2004), 1–76. (cited on pages 5, 27, 75, and 92)

Dunleavy, D., 2005. The image and the archive: A semiotic approach. In *Handbook of visual communication: theory, methods, and media*. Laurence Erlbaum Associates Publishers. (cited on page 78)

El Naqa, I. and Murphy, M. J., 2015. What is machine learning? In *Machine Learning in Radiation Oncology* (Eds. I. Editors El Naqa; R. LI; and M. J. Murphy), 3–11. Springer. (cited on page 89)

Evans, M. and Jack, R., 1984. *Sources of English Legal and Constitutional History*. Butterworths. (cited on page 29)

Faria, C.; Girardi, R.; and Novais, P., 2013. Analysing the problem and main approaches for ontology population. In *Information Technology: New Generations (ITNG), 2013 Tenth International Conference on*, 613–618. IEEE. (cited on page 36)

Feldman, R. and Sanger, J., 2007. *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press. (cited on pages 7 and 88)

Flood, M. D. and Goodenough, O. R., 2015. Contract as automaton: The computational representation of financial agreements. *Available at SSRN 2648460*, (2015). (cited on pages 44 and 45)

Foundation, W. W. W., 1991. History of the web. http://webfoundation.org/about/vision/history-of-the-web/. (cited on pages 72 and 73)

Fox, C., 2006. *Introduction to Software Engineering Design: Processes, Principles and Patterns with UML2*. Addison-Wesley Longman Publishing Co., Inc. (cited on pages 46, 47, and 50)

Francis, W. N. and Kucera, H., 1964. A Standard Corpus of Present-Day Edited American. Revised 1971, Revised and Amplified 1979. Department of Linguistics, Brown University Providence, Rhode Island, USA. http://www.hit.uib.no/icame/brown/bcm.html. (cited on page 86)

Freeman, M., 2001. *Lloyd's Introduction to Jurisprudence*. Sweet & Maxwell, London, 7th edn. (cited on pages 7, 13, 14, 16, and 17)

Friman, M. S., 1994. Plain english statutes-long overdue or underdone. *Loy. Consumer L. Rep.*, 7 (1994), 103. (cited on page 5)

GARDNER, M., 1970. Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223, 4 (1970), 120–123. (cited on page 53)

GLPI AND SMOLENKA, V., 2000. A Report on the Results of Usability Testing Research on Plain Language Draft Sections of the Employment Insurance Act. http://www.davidberman.com/wp-content/uploads/glpi-english.pdf. (cited on page 5)

GOLDSTEIN, J., 1999. Emergence as a construct: History and issues. *Emergence*, 1, 1 (1999), 49–72. (cited on page 56)

GOODALL, N., 2014. Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, , 2424 (2014), 58–65. (cited on page 57)

GORDON, T. F., 2008. Artificial intelligence and legal theory at law schools. *Artificial Intelligence and Legal Education*, (2008), 53–58. (cited on page 13)

GOVERNATORI, G., 2005. Representing business contracts in RuleML. *International Journal of Cooperative Information Systems*, 14, 2-3 (2005), 181–216. (cited on pages 7 and 45)

GRAMMEL, L. AND STOREY, M., 2010. Towards a foundation for information visualization engineering. *VizWeek 2010: The Role of Theory in Information Visualization*, (2010). (cited on page 76)

GRASSMANN, W. K. AND TREMBLAY, J. P., 1996. *Logic and discrete mathematics-a computer science perspective*. Prentice Hall. (cited on page 26)

GREENE, B., 2001. Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, 24, 1 (2001), 82–98. (cited on page 100)

GREENLEAF, G., 2010. The global development of free access to legal information. *European Journal of Law and Technology*, 1, 1 (2010). http://ejlt.org/article/view/17. (cited on pages 3 and 71)

GREENLEAF, G.; MOWBRAY, A.; AND CHUNG, P., 2013. The meaning of "free access to legal information": A twenty year evolution. *Journal of Open Access to Law*, 1, 1 (2013). (cited on page 3)

GREENLEAF, G.; MOWBRAY, A.; AND KING, G., 1995. Public legal information via internet: Austlii's first six months. *Australian Law Librarian*, 3 (1995), 144. (cited on page 74)

GRINSTEIN, G., 2010. Arguments for a theory of visualization. *VizWeek 2010: The Role of Theory in Information Visualization*, (2010). (cited on page 6)

GROSOF, B. N.; LABROU, Y.; AND CHAN, H. Y., 1999. A declarative approach to business rules in contracts: courteous logic programs in xml. In *Proceedings of the 1st ACM conference on Electronic commerce*, 68–77. ACM. (cited on page 45)

HAAPIO, H., 2011. Contract clarity through visualization–preliminary observations and experiments. In *Information Visualisation (IV), 2011 15th International Conference on*, 337–342. IEEE. (cited on page 106)

HAAPIO, H., 2013. *Next Generation Contracts: A Paradigm Shift*. Ph.D. thesis, University of Vaasa, Vaasa. (cited on pages 7, 21, 46, 47, 49, and 106)

HAAPIO, H. AND PASSERA, S., 2012. Reducing contract complexity through visualization-a multi-level challenge. In *Information Visualisation (IV), 2012 16th International Conference on*, 370–375. IEEE. (cited on page 106)

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; AND WITTEN, I. H., 2009. The WEKA Data Mining Software. *SIGKDD Explorations*, 11, 1 (2009). (cited on page 93)

HALVEY, M. AND KEANE, M., 2007. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, 1313–1314. ACM. (cited on page 7)

HART, H. L. A., 1961. *The concept of law*. Oxford University Press. (cited on pages 7, 13, 14, and 52)

HATHAWAY, O. A., 2001. Path dependence in the law: The course and pattern of legal change in a common law system. *The Iowa Law Review*, 86, 2 (2001). (cited on page 60)

HINSON, C. L., 2005. Legal informatics: Opportunities for information science. *Journal of Education for Library and Information Science*, 46, 2 (2005), pp. 134–153. http://www.jstor.org/stable/40323866. (cited on page 70)

HOLLAND, T. E., 1924. *The elements of jurisprudence*. Oxford Clarendon Press, 13th edn. (cited on page 41)

HOLZ, B., 2007. Chaos worth having: Irreducible complexity and pragmatic jurisprudence. *Minnesota Journal of Law Science and Technology*, 8 (2007), 303–715. (cited on pages 14 and 61)

HORSTMANN, C. S., 2008. *Big Java*. John Wiley & Sons, Inc. (cited on pages 43 and 50)

INDUKURI, K. V. AND KRISHNA, P. R., 2010. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference* (Bangalore, India, 2010), 1–5. New York, NY, USA. doi:http://doi.acm.org/10.1145/1754288.1754295. (cited on page 5)

JACKSON, B. S., 1995. *Making Sense in Law: Linguistic, Psychological, and Semiotic Perspectives*, vol. 4 of *Legal Semiotics Monographs*. Deborah Charles Publications. (cited on pages 25 and 28)

JOHANSEN, S. AND ROBBINS, R. A., 2015. Art-iculating the analysis: Systemizing the decision to use visuals as legal reasoning. *Legal Writing*, 20 (2015), 57–59. (cited on page 80)

JOHNSON, J., 2010. Embracing design in complexity. In *Embracing Complexity in Design* (Eds. J. J. KATERINA ALEXIOU AND T. ZAMENOPOULOS). Routledge. (cited on page 62)

JONES, B. C., 2013. Don't be silly: Lawmakers rarely read legislation and oftentimes don't understand it, but that's okay. *Penn State Law Review Penn Statim*, 118 (2013), 7. (cited on page 29)

JONES, G. T., 2008. Dynamical jurisprudence: law as a complex system. *Georgia State University Law Review*, 24, 4 (2008). (cited on page 58)

JURAFSKY, D. AND MARTIN, J. H., 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edn. (cited on page 27)

KAPITZE, C., 2009. Rethinking copyrights for the library through creative commons. *Library Trends*, 58, 1 (2009), 95–108. (cited on page 50)

KARLBERG, M. R., 2004. *Beyond the culture of contest: From adversarialism to mutualism in an age of interdependence*. George Ronald. (cited on page 17)

KIMBLE, J., 1994. Answering the critics of plain language. *Scribes J. Leg. Writing*, 5 (1994), 51. (cited on page 5)

KRIPS, H., 2013. Measurement in quantum theory. In *The Stanford Encyclopedia of Philosophy* (Ed. E. N. ZALTA). Stanford University, fall 2013 edn. (cited on page 68)

KRISHNAMURTHI, S., 2008. Teaching programming languages in a post-linnaean age. *ACM Sigplan Notices*, 43, 11 (2008), 81. (cited on page 43)

LANGTON, C. G., 1986. Studying artificial life with cellular automata. *Physica D: Nonlinear Phenomena*, 22, 1 (1986), 120–149. (cited on page 53)

LAURITSEN, M., 2007. Current frontiers in legal drafting systems. In *Proceedings of the 11th International Conference on AI and Law*. (cited on page 45)

LAVI, S., 2010. Turning the tables on "law and": A jurisprudential inquiry into contemporary legal theory. *Cornell Law Review*, 96, 4 (2010), 811–838. (cited on pages 7 and 20)

LEIBOFF, M. AND THOMAS, M. N., 2004. *Legal theories in principle*. Lawbook Co. (cited on pages 15 and 57)

LEITH, P., 2010. The rise and fall of the legal expert system. *European Journal of Law and Technology*, 1, 1 (2010). http://ejlt.org/article/view/14. (cited on page 71)

LENGLER, R. AND EPPLER, M., 2007. Towards a periodic table of visualization methods for management. In *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*. (cited on pages 6 and 76)

LETWIN, S. R., 2005. *On the History of the Idea of Law*. Cambridge University Press. (cited on pages 7, 14, and 16)

LOHMANN, S.; ZIEGLER, J.; AND TETZLAFF, L., 2009. Comparison of tag cloud layouts: Task-related performance and visual exploration. *Human-Computer Interaction–INTERACT 2009*, (2009), 392–404. (cited on page 7)

LOPUCKI, L. M., 1996. Systems approach to law. *Cornell Law Review*, 82 (1996), 479. (cited on pages 61 and 62)

MANDERSON, D., 2012. The law of the image and the image of the law: Colonial representations of the rule of law. *New York Law School Law Review*, 57 (2012), 153–205. (cited on page 22)

MARTIN, P. The mushrooming virtual law library on the net. In *Cornell Law Forum*, vol. 27. (cited on page 3)

MARTIN, P., 2010. How structural features of the U.S. judicial system have affected the take-up of digital technology by courts. *European Journal of Law and Technology*, 1, 1 (2010). http://ejlt.org/article/view/16. (cited on page 71)

MARTIN, P. W., 1999. Internet: Full and unfettered access to law-some implications, the. *N. Ky. L. Rev.*, 26 (1999), 181. (cited on page 3)

McEnery, T., 2005. Corpus linguistics. In *The Oxford handbook of computational linguistics* (Ed. R. Mitkov). Oxford University Press. (cited on page 89)

Melham, D., 1993. Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs. (cited on page 5)

Mellinkoff, D., 1963. *The Language of the Law*. Little, Brown and Company. (cited on pages 4 and 29)

Mertz, E. and Rajah, J., 2014. Language-and-law scholarship: An interdisciplinary conversation and a post-9/11 example. *Annual Review of Law and Social Science*, 10 (2014), 169–183. (cited on pages 7 and 20)

Mingers, J., 1980. Towards an appropriate social theory for applied systems thinking: Critical theory and soft systems methodology. *Journal of Applied Systems Analysis*, 7 (1980), 41–50. (cited on page 64)

Molina-Jimenez, C.; Shrivastava, S.; Solaiman, E.; and Warne, J., 2004. Run-time monitoring and enforcement of electronic contracts. *Electronic Commerce Research and Applications*, 3, 2 (2004), 108–125. (cited on pages 44 and 45)

Moriarty, S., 2005. Visual semiotic theory. In *Handbook of visual communication: theory, methods, and media* (Eds. K. Smith; S. Moriarty; G. Barbatsis; and K. Kenney), LEA's Communication Series. Laurence Erlbaum Associates Publishers. (cited on pages 77 and 78)

Moriarty, S. and Sayre, S., 2005. An intended-perceived study using visual semiotics. In *Handbook of visual communication: theory, methods, and media*. Laurence Erlbaum Associates Publishers. (cited on page 78)

Moriarty, S. E. and Kenney, K., 1995. Visual communication: A taxonomy and bibliography. *Journal of Visual Literacy*, 15, 2 (1995), 7–156. (cited on page 76)

Nicolis, G. and Nicolis, C., 2007. Foundations of complex systems: nonlinear dynamic statistical physics information and prediction. *Foundations of complex systems: nonlinear dynamic statistical physics information and prediction*, (2007). (cited on pages 58 and 64)

Norman, D., 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books. (cited on page 46)

NZ, P., 2008. Presentation of New Zealand Statute Law. Technical Report 104, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office. (cited on page 74)

OLIVER, L., 2002. *The beginnings of English law*, vol. 14. University of Toronto Press. (cited on page 24)

PALIWALA, A., 2010. A history of legal informatics. *European Journal of Law and Technology*, 1, 1 (2010). http://ejlt.org/article/view/21. (cited on pages 70 and 72)

PASSERA, S., 2012. Enhancing contract usability and user experience through visualization-an experimental evaluation. In *Information Visualisation (IV), 2012 16th International Conference on*, 376–382. IEEE. (cited on page 106)

PASSERA, S.; HAAPIO, H.; AND CURTOTTI, M., 2014. Making the meaning of contracts visible - automatic contract visualization. In *Transparency. Proceedings of the 17th International Legal Informatics Symposium IRIS 2014. Jusletter IT*. Österreichische Computer Gesellschaft OCG, Available at SSRN: http://ssrn.com/abstract=2630609. (cited on pages xiii, 12, 22, 87, 88, 106, and 107)

PENNER, J. E.; SCHIFF, D.; NOBLES, R.; AND BARRON, A. E., 2002. *Introduction to jurisprudence and legal theory: commentary and materials*. Oxford University Press. (cited on pages 7, 13, and 17)

PORTER, E. G., 2014. Taking images seriously. *Columbia Law Review*, (2014), 1687–1782. (cited on page 79)

R-CORE-TEAM ET AL., 2012. R: A language and environment for statistical computing. (2012). (cited on page 93)

RAY, C., 2000. Logical positivism. *A Companion to the Philosophy of Science*, (2000), 243–251. (cited on page 66)

REINGOLD, B., 1996. Conflict and cooperation: Legislative strategies and concepts of power among female and male state legislators. *The Journal of Politics*, 58, 02 (1996), 464–485. (cited on page 17)

RENTON, D., 1975. The preparation of legislation - report of a committee appointed by the Lord President of Council. Technical report, Council of the UK Government. (cited on page 29)

ROSMAN, A. L., 2013. Visualizing the law: Using charts, diagrams, and other images to improve legal briefs. *Journal of Legal Education*, 63 (2013), 70. (cited on page 79)

RUHL, J., 2008. Law's complexity-a primer. In *Georgia State University Law Review Symposium Issue, Forthcoming*. (cited on pages 58 and 59)

RUHL, J. AND SALZMAN, J., 2002. Mozart and the red queen: The problem of regulatory accretion in the administrative state. *Geo. LJ*, 91 (2002), 757. (cited on pages 57, 59, and 60)

RUSSELL, S. AND NORVIG, P., 2003. *Artificial Intelligence: A Modern Approach, 2003*. Prentice Hall, Pearson Education Inc, 2nd edn. (cited on pages 26 and 89)

SANTOS, L. E.; GIRARDI, R.; AND NOVAIS, P., 2013. A case study on the construction of application ontologies. In *Information Technology: New Generations (ITNG), 2013 Tenth International Conference on*, 619–624. IEEE. (cited on page 36)

SAWYER, A., 2010. Enhancing compliance through improved readability: Evidence from New Zealand's rewrite "experiment". *Recent Research on Tax Administration and Compliance*, (2010). (cited on page 5)

SCHULMAN, L. AND SEIDEN, P., 1978. Statistical mechanics of a dynamical system based on Conway's Game of Life. *Journal of Statistical Physics*, 19, 3 (1978), 293–314. (cited on page 53)

SEIPEL, P., 2004. IT law in the framework of legal informatics. *Scandinavian Studies in Law*, 48 (2004), 32–47. (cited on page 70)

SHANNON, C. E., 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (1948), 379–423. (cited on page 33)

SIPPER, M., 1995. An introduction to artificial life. *Explorations in Artificial Life (special issue of AI Expert)*, (1995), 4–8. (cited on page 53)

SISSON RUNYAN, A., 1994. Feminism: Alternative futures. *In Beckman, Peter R and D'Amico, Francine. (eds). Women, gender, and world politics: Perspectives, policies, and prospects*, (1994). (cited on page 17)

SMITH, A. R., 1984. Plants, fractals, and formal languages. *ACM SIGGRAPH Computer Graphics*, 18, 3 (1984), 1–10. (cited on page 53)

SMITH, D. AND RICHARDSON, G., 1999. The readability of Australia's taxation laws and supplementary materials: an empirical investigation. *Fiscal Studies*, 20, 3 (1999), 321–349. (cited on page 5)

Smith, K.; Moriarty, S.; Barbatsis, G.; and Kennedy, K., 2005. *Handbook of visual communication: theory, methods, and media*. LEA's Communication Series. Lawrence Erlbaum Associates Publishers. (cited on page 76)

Smits, J. M., 2014. Law and interdisciplinarity: On the inevitable normativity of legal studies. *Critical Analysis of Law*, 1, 1 (2014), 75–86. (cited on page 69)

Spulber, D. F. and Yoo, C. S., 2004. On the regulation of networks as complex systems: A graph theory approach. *Northwestern University Law Review*, 99 (2004), 1687. (cited on page 42)

St Amant, K., 2007. *Handbook of Research on Open Source Software: Technological, Economic, and Social Perspectives: Technological, Economic, and Social Perspectives*. IGI Global. (cited on page 73)

Stefanou, C. and Xanthaki, H., 2013. *Drafting legislation: a modern approach*. Ashgate Publishing, Ltd. (cited on page 28)

Summers, R. S. and Howard, C. G., 1972. *Law: Its nature, functions, and limits*. Prentice-Hall, 2nd edn. (cited on pages 51 and 52)

Surden, H., 2012. Computable contracts. *UCDL Rev.*, 46 (2012), 629. (cited on pages 30, 32, and 44)

Swienty, O. and Takatsuka, M., 2010. From subjectivity to objectivity in information visualization: challenges in bridging the gap between cognitive theory and information design. *VizWeek 2010: The Role of Theory in Information Visualization*, (2010). (cited on pages 6 and 76)

Tanner, E., 2002. Seventeen years on: Is Victorian legislation less grammatically complicated. *Monash UL Rev.*, 28 (2002), 403. (cited on page 5)

Thompson, S., 1999. *Haskell: the craft of functional programming*. Addison-Wesley, 2nd edn. (cited on page 43)

Tickner, J. A., 1994. A feminist critique of political realism. *In Beckman, Peter R and D'Amico, Francine. (eds). Women, gender, and world politics: Perspectives, policies, and prospects*, (1994). (cited on page 17)

Tomlins, C. and Comaroff, J., 2011. 'law as...': Theory and practice in legal history. *Theory and Practice in Legal History*, (2011), 2012–50. (cited on pages 7 and 20)

TWINING, W., 2001. Some jobs for jurisprudence (1974). *Lloyd's Introduction to Jurisprudence, 7th edition, London: Sweet & Maxwell*, (2001), 21–24. (cited on page 11)

VAN ROY, P. ET AL., 2009. Programming paradigms for dummies: What every programmer should know. *New computational paradigms for computer music*, 104 (2009). (cited on page 43)

VEITCH, S.; CHRISTODOULIDIS, E.; AND FARMER, L., 2007. *Jurisprudence: themes and concepts*. Routledge-Cavendish. (cited on page 57)

VENTURI, G., 2008. 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*, 1. (cited on page 112)

WACKS, R., 1999. *Jurisprudence*. Blackstone Press Limited, fifth edn. (cited on pages 7, 13, 15, and 16)

WAGNER, G., 1986. Interpreting cloze scores in the assessment of text readability and reading comprehension. (cited on page 100)

WEBBER, J., 2009. The grammar of customary law. *McGill Law Journal*, 54 (2009), 579. (cited on pages 41 and 42)

WILLIAM, J., 1980. *Legal Philosophies*. Butterworths & Co Publishers Ltd. London. (cited on pages 7, 13, 14, and 15)

WINKELS, R. AND HOEKSTRA, R., 2013. Automatic extraction of legal concepts and definitions. In *Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference*, vol. 250, 157. IOS Press. (cited on page 7)

WOLFRUM, R., 1983. The principle of the common heritage of mankind. *Heidleberg Journal of International Law*, 43 (1983), 312. (cited on page 73)

WOODS, B.; MOSCARDO, G.; GREENWOOD, T.; ET AL., 1998. A critical review of readability and comprehensibility tests. *Journal of Tourism Studies*, 9, 2 (1998), 49–61. (cited on pages 5 and 75)

YI, J., 2010. Toward practical information visualization theory: Lessons from the history of decision science. *VizWeek 2010: The Role of Theory in Information Visualization*, (2010). (cited on page 76)

Zins, C., 2007. Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology*, 58, 4 (2007), 479–493. (cited on pages 33, 34, and 76)

# Machine Learning for Readability of Legislative Sentences

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

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

Tom Bruce
Legal Information Institute
Cornell University Law School
Ithaca, NY
tom@liicornell.org

Sara Frug
Legal Information Institute
Cornell University Law School
Ithaca, NY
sara@liicornell.org

Wayne Weibel
Legal Information Institute
Cornell University Law School
Ithaca, NY
wayne.weibel@liicornell.org

Nicolas Ceynowa
Legal Information Institute
Cornell University Law School
Ithaca, NY
nic.ceynowa@liicornell.org

## ABSTRACT

Improving the readability of legislation is an important and unresolved problem. Recently, researchers have begun to apply legal informatics to this problem. This paper applies machine learning to predict the readability of sentences from legislation and regulations. A corpus of sentences from the United States Code and US Code of Federal Regulations was created. Each sentence was labelled for language difficulty using results from a large-scale crowdsourced study undertaken during 2014. The corpus was used as training and test data for machine learning. The corpus includes a version tagged using the Stanford parser context free grammar and a version tagged using the Stanford dependency grammar parser. The corpus is described and made available to interested researchers. We investigated whether extending natural language features available as input to machine learning improves the accuracy of prediction. Among features evaluated are those from the context free and dependency grammars. Letter and word ngrams were also studied. We found the addition of such features improves accuracy of prediction on legal language. We also undertake a correlation study of natural language features and language difficulty drawing insights as to the characteristics that may make legal language more difficult. These insights, and those from machine learning, enable us to describe a system for reducing legal language difficulty and to identify a number of suggested heuristics for improving the writing of legislation and regulations.

## Keywords

readability, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, plain language, supervised learning, legislative drafting

## 1. INTRODUCTION

A standing problem of legal language is its low readability. Legislation, the focus of our study, is now read by a broad audience who access it online. Further, readability research on regulatory texts suggests that legislative language is difficult to entirely inaccessible for many of its readers, even after redrafting according to plain language principles. Our previous research suggests the legally trained may now be a minority of those who read legislation online. Traditional readability metrics have been seen as of limited value for the prediction of the difficulty of legal language. The research to date establishes both an audience for, and a need to improve the readability of legal language.(Curtotti et al., 2015)

In this study, we apply legal informatics to this problem. We prepare and make available a corpus of legislative sentences labelled for language difficulty. The corpus is provided in plain language and grammatically marked up versions (both context free grammar and dependency grammar). We investigate machine learning to predict the readability of legal sentences using both fully parsed and unparsed natural language features including letter and word ngrams and a variety of features derived from the parsed versions of the corpus. Visualization tools and correlation analysis are used to provide insights into characteristics of legislative language that contribute to its difficulty. These insights enable us to suggest a system for assisting writers of legal language and heuristics for legislative drafting.

## 2. RELATED WORK

In a previous phase of our work, we reported research on the readability of legal language.(Curtotti et al., 2015) Over a three month period in 2014, 43,000 crowd sourced assessments were collected on the difficulty of legal and other language. The data was collected from users of the Cornell Legal Information Institute online legislative pages.[1] Users were asked to rate 1255 sentences for reading difficulty, including approximately 500 sentences from the United States Code and 500 sentences from the US Code of Federal Regulations. As far as we are aware, this was the largest ever crowdsourced collection of user assessments of the difficulty of legislative language. A raw reading difficulty score and

---

[1] http://www.law.cornell.edu/

"easy" or "hard" labels were assigned to the test sentences based on collation and analysis of responses submitted by users. The study also collected demographic data provided by the users of online legislation. Further we carried out initial machine learning with limited features on the task of predicting the easy/hard classification assigned to the test sentences. In this phase, a limited number of natural language features were extracted. A support vector machine algorithm within the Weka machine learning software was used to predict whether sentences were easy or hard.(Hall et al., 2009)

Findings and results reported included the following:

(a) a very small proportion of the US Code is read very often, while the bulk of the Code is read very rarely;

(b) legal professionals (including law students) were a minority among research respondents (a result suggesting that legal professionals may now be a minority of readers of regulatory materials);

(c) women, those without tertiary education and Spanish speakers are proportionally under-represented among research participants;

(d) law is easier to read for legal professionals (including law students) than for other users of online legal rules;

(e) we demonstrated the feasibility of long-term collection of online assessment of the readability and usability of legal texts using crowdsourced methods;

(f) from the data collected it was possible to establish a ranking of legislative sentences by language difficulty; and

(g) initial investigation of machine learning algorithms established that they outperform the accuracy of traditional readability metrics in predicting the readability of legislative texts.

The data we collected in our previous research is further investigated here. Other related research is extensively reviewed in our previous paper. That review included readability studies, plain language, readability and legislation, crowdsourced assessment of language difficulty, natural language processing and machine learning, sentence level assessment of language difficulty, likert testing, cloze testing, semantic differentials and principal components analysis. Rather than reproducing that related research here, readers are referred to that paper.(Curtotti et al., 2015)

Below we highlight some recent research particularly relevant to the work undertaken in this paper; focussing on machine learning and the exploitation of the graph characteristics of language.

Recent years have seen research applying natural language processing and machine learning to assessing the readability of natural language. Most recently, a number of researchers have shifted their focus from assessing the readability of documents to assessing the readability of sentences. This latter focus is of particular interest in the context of our own research which is also focussed at sentence level.

Machine learning for readability requires input data labelled for readability. Labelling data is the most resource intensive part of machine learning as it is typically done by human judges. A dataset for machine learning can be expressed as a matrix of features. Each row of the matrix represents the features of a data instance (in our case a sentence). Ideally each row of the matrix is also labelled with its appropriate classification. The task of machine learning (when a classification label is available) is to predict the label of each row of data. In application to readability, machine learning can be used to develop a model to predict the readability classification of a new input given prior training of the machine learning algorithm using the labelled training data. A variety of algorithms have been developed to undertake such prediction tasks. In our research, including for this paper, we have found that a support vector machine (SVM) generally outperforms other machine learning algorithms in predicting readability. This is the algorithm we have focussed on for this research.[2](Curtotti et al., 2015; Curtotti and McCreath, 2013)

In 2014, Dell'Orletta et al. addressed the question of predicting the readability of sentences. They observe that while currently text difficulty is primarily assessed at document level, simplification is carried out at sentence level. This makes it difficult to assess whether possible text simplification addresses actual readability problems in the text. They were particularly interested to better understand what kinds of corpora (training/test data) and what kinds of input features, had the greatest positive impact on accuracy of prediction. In respect of input features, they found, as compared with document level classification, sentence level readability classification requires a large number of input features - mainly syntactic ones. Among the features they examine are raw text features (e.g. sentence length and word length), lexical features (vocabulary, type to token ratio), morpho-syntactic features (parts of speech, verbal characteristics) and syntactic features (parse tree depths, conjunctions, chains of subordinate clauses, length of dependency links). They conclude that raw text features, syntactic and morpho-syntactic features all contribute to "achieving adequate performance". Their results also suggest that the task of predicting sentence difficulty is more complex than for document readability. Only a small number of features (circa 10) were required for high accuracy on documents, whereas they showed accuracy improved gradually as features were increased to 66 for sentences. They compared machine learning performance on different corpora by varying the 'difficult language' corpus. Either they used the entire corpus (which contained some easy sentences), or manually cleaned the difficult language corpus to remove easy sentences. They did not achieve a notable improvement in accuracy by this latter process, suggesting that larger noisy corpora are effective as input for readability prediction.(Dell'Orletta et al., 2014)

Falkenjack and Johnsson examine whether using a small set

---

[2]For example on the readability tasks we investigate, an SVM outperforms algorithms such as rule based algorithms (e.g. JRip), decision trees (J48 tree) or Naive Bayes

of features, not requiring syntactic parsing of natural language, achieves similar accuracy to that achieved by using a larger feature set. They used a genetic algorithm to select the best performing features. The feature set they explore is similar to that used by Dell'Orletta et al. Among the parse based features they examine are average dependency distance on a document and per sentence level, ratio of right dependencies, average sentence depth and ratio of dependency types. They show that a small number of features can attain a high accuracy.(Falkenjack and Jönsson, 2014) However their work is carried out at document level and is not consistent with the results of Dell'Orletta et al at sentence level.

Both papers cited above include dependency grammar features in their feature set. Dependency grammars represent language as lexical items (e.g. pairs of words) linked by asymmetrical dependency relationships. One word will be a head and another a dependent.(Nivre, 2005) For example, in the phrase 'the quick brown fox ran', 'the' is dependent on 'fox', while 'fox' is dependent on 'ran'. The Stanford dependency grammar is an example of a dependency grammar. The Stanford grammar (like other dependency grammars) does not represent phrase structures but rather represents grammatical relations as typed triples in which there is a head word and dependent word. The dependency grammar of a sentence constitutes a network that can be represented as a directed graph.(De Marneffe and Manning, 2008)

Cong and Liu undertake a review of the application of complex networks to human language. Such networks can be used to represent language and provide access to quantitative measures and a model which corresponds to the nature of language as a set of relations between linguistic units (i.e. a network). A network is comprised of a set $N = (V, E)$ where V is a set of vertices and E is a set of edges representing linkages between vertices. Cong and Liu distinguish between static linguistic networks and dynamic linguistic networks. An example of the former is the semantic network formed by the lexicon of a language through relationships of meaning.(e.g. relationships of synonymy or antonymy). Dynamic linguistic networks arise from actual natural language use and can represent various aspects of language 'along the meaning form dimension'. Thus a network can be built from the linear order of words (forming a word co-occurrence network), or a network can be formed from the syntactic or semantic dependency between words. A number of quantitative measures are used to characterise the shape (topology) of graphs. Among such measures are density (proportion of edges to all possible edges), degree distribution (probability of a vertex having degree (k)), average shortest path length connecting vertices, clustering coefficient and network centralisation. Such measures have been used to characterise language as a multilevel system; to classify natural language into different languages using network parameters;(Liu and Cong, 2013),(Liu and Li, 2010) and in application to the micro-characteristics of language.(Cong and Liu, 2014)

## 3. A LABELLED AND PARSED CORPUS OF AMERICAN REGULATORY ENGLISH

We have created a corpus which is available on request from the authors. The corpus consists of four sets of natural language in three parallel corpora, as well as other supporting files.

The underlying natural language is a set of randomly extracted sentences from four bodies of natural language as follows:

(a) 466 sentences randomly drawn from the United States Legal Code downloaded from the Legal Information Institute.

(b) 424 sentences randomly drawn from the United States Code of Federal Regulations, also downloaded from the Legal Information Institute.[3]

(c) 117 sentences randomly drawn from the Brown Corpus of American written English. This corpus represents 'normal' English usage.(Francis and Kucera, 1964)

(d) 134 sentences randomly drawn from a corpus of graded readers which are graded into six reading levels.[4]

The non-legal sentences included in the corpus were used for calibration purposes in our study and are useful in interpreting the results of analyses carried out on legal sentences. They are included in the corpus for this purpose. As sentences were automatically extracted, some errors occurred in the segmentation process resulting in inclusion of material that did not constitute sentences. Manual cleaning was undertaken of the entire corpus, to remove such material, resulting in a slightly smaller corpus than that used in our previous study.

The corpus is also provided in three parallel versions:

(1) csv files containing original texts, related markup and for the legal sentences, original html markup is preserved;

(2) context free grammar parsed versions of the sentences produced using the Stanford Parser;(Klein and Manning, 2003) and

(3) dependency grammar representations of the sentences produced using the Stanford Parser.(De Marneffe et al., 2006)

To facilitate use, the corpus is packaged with a python script which allows the data to be accessed in raw, labelled and parsed forms, using a version of python in which the Natural Language Toolkit is installed.(Bird et al., 2009) The dependency grammar representations of sentences can be accessed as mathematical graphs using networkx (a python library for network analysis).

Although the corpus is small in the scheme of things, its availability as readability labelled data makes it a valuable

---

[3]Legal Information Institute, Cornell University Law School, http://lii.org.
[4]Graded sentences were extracted from graded reader passages downloaded from http://www.lextutor.ca/graded/. These passages are no longer available at time of publication. A copy of the corpus can be obtained for research purposes by contacting the authors.

resource for studying the readability of American regulatory english. It may be noted that as the language included in the corpus is all drawn from the US Federal jurisdiction, it represents a subset of regulatory english. It is likely to be a suitable body of materials within the broader US regulatory context, but it would not necessarily be valid to extend conclusions to non-US jurisdictions where regulatory language is different.

# 4.   MACHINE LEARNING FOR PREDICTING AND IMPROVING THE READABILITY OF REGULATORY ENGLISH

Collecting user evaluations of sentences is time consuming and difficult. It requires the availability of online infrastructure and access to audience. It calls on the time of users who are asked to provide evaluations. As our previous study demonstrated, large-scale collection of such user evaluations is feasible and can provide valuable insight into the readability of legal language. Ideally, we would wish to be able to predict readability of a sentence without having to conduct such surveys.

To explore the application of machine learning to our dataset we divided the sentences into two approximately equally sized "easy" or "hard" classes. The class assigned depended on the language difficulty of the sentence, measured using a numerical language difficulty score. This score was created by combining responses provided by multiple users to different tests on each sentence. The measure is further described in our previous paper.(Curtotti et al., 2015) In that research we undertook machine learning after extraction of a limited number of features from the test sentences. These features included: sentence length; average word length; type to token ratio (i.e. ratio of unique words to total words); common readability metrics; proportion of verbal phrase chunks; and proportional distribution of different parts of speech.

In our previous study, an F-measure accuracy of 72.7% was attained in predicting the easy/hard classification on data from all four corpora together. We found that readability metrics made little contribution to accuracy and were of little value in predicting the readability of sentences in a machine learning framework. We also carried out machine learning on purely legal sentences attaining an accuracy of 70.5%, i.e. a little less than for the four corpora dataset. This indicates the greater difficulty of learning a model to distinguish easy from hard legal sentences. For the legal dataset, if only readability metrics were used as input, accuracy was 60.2%.

To further investigate the relationship between language difficulty and readability metrics we also examined correlation between readability metrics and the numerical language difficulty score. The SMOG index was most highly correlated at 0.33, which was about the same as for sentence length. This was however exceeded by the type to token ratio at -0.42. In other words, it is more effective to count the ratio of unique words to total words as a measure of language difficulty than to rely on readability metrics at sentence level. We also concluded that natural language processing derived features are more effective than readability metrics as input for machine learning for readability both in the case of the purely legal corpora and in the case of all four corpora together. These results showed machine learning can be used as a better predictor of readability of legal sentences than traditional readability metrics.

In this paper we extend our exploration of machine learning on this dataset. In particular, in the context of regulatory language:

(1) Does adding ngram features improve accuracy in predicting reading difficulty classification?

(2) Does adding parsed features from context free grammar and dependency grammar derived features including graph characteristics (such as those discussed by Liu et al. and used in the Falkenjack et al. and Dell'Orletta et al. papers), increase prediction accuracy?

(3) Does increasing the overall number of available features for machine learning at sentence level increase accuracy?

To address these questions we first extracted ngram features using the Open Online Platform for Readability Research. (Curtotti and McCreath, 2013) Ngrams are multiples of linguistic units. For example, a sequence of three letters is a 3-gram of letters, two sequential words are a 2-gram of words. For the research reported here we extracted both letter grams (1, 2, 3 and 4 grams) and word grams (1, 2 and 3 grams). As the number of ngrams for our corpus was in the thousands and made machine learning impractical with the software available to us, a filter was applied to limit ngrams extracted to those which occurred at least 30 or 50 times in the corpus, depending on the ngram level. Letter ngrams potentially capture morpho-syntactic features of language (e.g. the letter combination 'ment' often represents a noun, while the letter combination 'ly' often occurs in adverbs). Word ngrams potentially capture higher level units of linguistic meaning. For example, the word combination 'intellectual property' has a distinctive legal meaning. Further, word ngrams can also stand in for grammatical constructs. For example the occurrence of a preposition (in, on, before etc) may mark the beginning of a subordinate prepositional phrase. An advantage of use of ngrams over other features, is that they do not require full grammatical parsing of language, as ngrams are simply extracted by counting token occurrences. If such features could stand in place of syntacticly parsed features, it would simplify the task of sentence classification.

We also extracted context free grammars (CFG) and dependency grammar parsed versions of the input sentences using the Stanford parser.(Klein and Manning, 2003; De Marneffe et al., 2006) Both versions were extracted in raw counts (i.e. results were not normed for sentence length). The CFG parses were represented as counts of particular grammatical types at word and phrase level. The dependency grammar was similarly extracted and represented. In addition, some topological graph characteristics were extracted from both parses and used as input to machine learning. Features connected with these parses included:

(a) CFG parse tree height and number of sub-trees (as measures of syntactic complexity);

(b) CFG tree slope (this marks where in a sentence syntactic complexity is found);

(c) dependency grammar average clustering;

(d) dependency grammar clique average;

(e) average degree;

(f) counts of degree up to degree fifteen;

(g) number of edges and number of nodes;

(h) mean dependency distance;

(i) ratio of past tense verbs to all verbs; and

(j) ratio of content (nominal) words

A limitation of both the Stanford parsed dependency parse and the Stanford context free grammar parse was that for very long sentences parsing failed in both cases. This would result in a generic dependency or CFG class being assigned by the parser to such cases. This data was included in machine learning. While the features for these very long sentences are not terribly helpful and will have affected the result, removing these instances would mean the dataset would no longer be a random sample from the US Code and Code of Federal Regulations. Such instances would for example affect machine learning on features such as sentence length and are therefore retained, even though parsing was partial.

As with our previous study, we used a Support Vector Machine (SMO) algorithm for machine learning, with default Weka settings. All reported results are on the basis of ten-fold cross validation. Where experiments involved a large numbers of features, supervised filtering of features was carried out to identify the best features for machine learning. The algorithm used for this purpose was Correlation Feature Subset Selection which selects the features most highly correlated with the prediction class while filtering attributes which are highly correlated with each other.(Hall, 1999)

## 4.1 Baseline Accuracy

As further data cleaning had been carried out since our earlier reported research, we re-tested baseline accuracy with features used in our previous paper to test whether these results remained valid. For all four corpora using features from our previous study accuracy was 72.9% (i.e. similar to our previous results). We also reassessed accuracy with subsets of earlier tested features. For readability metrics alone, accuracy was 67.7%. For surface features only (e.g. sentence length, average word length), accuracy was 69.6%. On the legal corpora only, accuracy for all features from our previous study was 70.8%, for readability metrics it was 63.8% and for surface features only 67.1%. Below we discuss the effect of addition of new features for machine learning. Results are summarised in Table 1.

## 4.2 Letter-grams and other features

A feature set was prepared initially containing 1713 letter ngrams and other features extracted from the data. This set was filtered to remove 'useless features' (resulting in the removal of 100 features). A second filter was applied using supervised learning to automatically select the best features for machine learning. The filter used was the Correlation-based Feature Subset Selection with Best-First selection using bi-directional best first search with search limited to a depth of 10. This resulted in a reduced feature set of 133 features. An accuracy of 74.4% was attained on all four corpora. Although the letter grams constituted the bulk of features selected by the search algorithm, features were also drawn from other feature sets: CFG grammar; NLTK normed parses; dependency grammar features and one readability metric (the Dale-Chall readability metric). It may be noted (although the difference is unlikely to be significant) that this feature set was less accurate than manually selected StanPhrases and CFG grammar graph features which resulted in an accuracy of 75.1%.

Also accuracy was trialled with ngram, traditional readability metrics and surface features only. Such features can be extracted without complex parsing being required and as suggested by Falkenjack et al. above, might be preferable for that reason. Again automatic feature selection was used and resulted in a reduced set of 110 features. In this trial, accuracy dropped to 72.1%. The same tests were carried out on the legal corpora only. In this case, automatic feature selection produced a feature set of 101 features. Accuracy was 72.1%. In the case where only ngram, surface and readability metric features were used as input, 97 features were automatically selected and accuracy was 71.6%. In this case, the difference in accuracy between use of parsed features and non-parsed features was marginal.

After addition of ngrams, accuracy on learning on all corpora and the two legal corpora increased (72.9% vs 74.4% for all corpora, and 70.8% to 72.1% for the legal corpora). However, as noted above, an even higher level of accuracy (i.e. 75.1%) was obtained using just the raw phrase and grammatical class counts extracted by the Stanford parser. The NLTK parse used in our first study used normed rather than raw counts (removing the effect of sentence length). However, our results show that this detracts significantly from the effectiveness of phrase features for machine learning for readability (assuming that there are no other significantly contributing differences between the NLTK and the Stanford parser parses). Further, counts of grammatical phrases and grammatical classes are significantly more effective in assessing language difficulty than simply counting the number of words in a sentence or using word length. Using only sentence length for prediction reduces accuracy to 57.4% on the legal database. With the addition of average word length, accuracy increases to 60%, still well short of the results achieved by parse features. This is notable, as sentence length and average word length are the primary features used in traditional readability metrics to measure language difficulty.

## 4.3 Word ngrams and other features

In the case of word ngrams we started with 729 learning features. After removing useless features, 709 features re-

mained. From this reduced set, the best features were automatically selected, leaving 86 learning features. Of these, 29 were word grams and the remainder were a selection of features from the remaining feature sets. On all four corpora an accuracy of 76.7% was attained. Using just ngram features, surface features and readability metrics resulted in an accuracy of 73.6%. In the case of the legal only corpora, after removal of useless features, 694 features remained. After automatic selection of best features, the feature set was reduced to 61 features. Of these, 23 were ngram features. Accuracy of machine learning was 73.8%. Using just non-parse features, as above, resulted in an accuracy of 70.7%.

Like letter ngrams, the presence of word ngrams increases accuracy. In this instance, the inclusion of word ngrams in the feature set, when combined with parse and other features, achieved the highest score of the various feature sets trialled during the study (i.e. 76.7% for all corpora, and 73.8% for the legal corpora).

We may conclude that adding ngrams increases accuracy, particularly in the case of word ngrams. Further adding complex syntactic derived features also increases accuracy as compared to non-parse features. Highest results are obtained by drawing on all available features. These results are consistent with previous research applied to non-legal sentences which pointed to the complexity of predicting readability at sentence level.(Dell'Orletta et al., 2014) Our investigation did not find that dependency derived features 'stand out' from other features as particularly useful in predicting readability, although they are included among automatically selected features.[5]

| Results | All Corps. | Legal |
| --- | --- | --- |
| Baseline - metrics only | 67.7% | 63.8% |
| Baseline - surface features only | 69.6% | 67.1% |
| Baseline - all features | 72.9% | 70.8% |
| New study: syntax features only | 75.1% | - |
| New study: non-parse + letter ngrams | 72.1% | 71.6% |
| New study: other + letter ngrams | 74.4% | 72.1% |
| New study: non-parse + word ngrams | 73.6% | 70.7% |
| New study: other + word ngrams | 76.7% | 73.8% |

**Table 1: Summary of Key Results**

# 5. DEVELOPING HEURISTICS FOR WRITING READABLE REGULATORY TEXTS

The machine learning results above represent a classification model which could be used as a plug-in for software to detect easy or hard legal sentences. However such a model, while useful, is not necessarily sufficient to assist a writer trying to improve readability. Simply knowing sentence is likely to be experienced by readers as hard, does not necessarily provide insight as to what the writer might ideally do to improve readability. Further, there is some artificiality implicit in applying machine learning to readability. Machine learning separates data into researcher assigned classes or can be used to detect natural clustering in a dataset. This is not ideally suited to readability, as language is not neatly partitioned into 'easy' and 'hard' classes. Rather language difficulty is

[5]As we see below in Table 2, they tend to have a lower correlation with readability than context free grammar derived features.

a continuum. This is illustrated by the distribution of the language difficulty scores for each sentence from which the easy, hard classification was derived. Figure 1 shows the density distribution (y-axis) of difficulty scores (x-axis) for each of the four corpora. Graded sentences are weighted to the easier end, but nonetheless the difficulty of sentences as a whole approximates a gaussian. Most of the data is distributed close to a central (artificially selected) dividing line between easy and hard sentences.
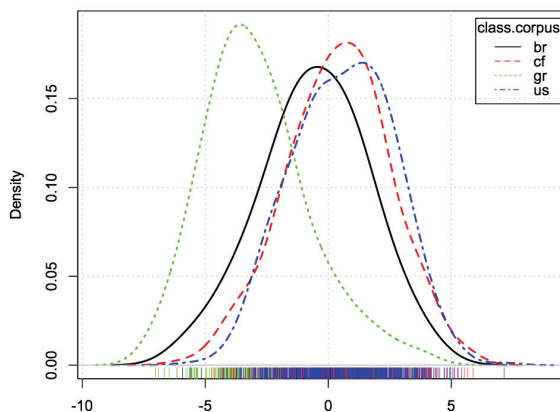


**Figure 1: Density distribution of language difficulty by corpus. gr = graded; br = brown; us = United States Code; cf = US Code of Federal Regulations**

This is not to suggest that machine learning is without value. It is useful, with some probability, to be able to say that a particular sentence is 'easy' or 'hard'. In our case such predictions can be made with around 74% accuracy on US regulatory sentences. Knowledge that a sentence is likely to be hard for readers serves as a focus for attention. The writer can then use their own experience as well as guidelines, such as those discussed in this article, to improve the sentence. While some false negatives will be ignored, overall a document should improve in readability through application of machine learning results.

Further the machine learning results provide insights that allow us to begin to develop and suggest heuristics specifically suited to the writing of regulatory texts (at least in the context of the US style of regulatory writing).

We have seen above that particular features can be extracted which are particularly helpful in carrying out machine learning. These features also tell us something about what contributes to language difficulty. Using the visualisation features provided by Weka we can see how the presence of particular features is related to language difficulty. Figure 2 provides an illustration. Here the image shows histogram distributions of the occurrence counts for a particular feature. Colour coding shows classification. Red represents sentences classified as hard. Blue represents sentences classified as easy. The top graph (A) shows the progressive increase in proportion of difficult sentences as the number of prepositions in a sentence increases. Most sentences having more than 6 prepositions are classified as hard. The bottom graph (B) shows occurrences of cardinal numbers. In legal

language these are often associated with cross references to other sections. Most sentences having 3 or more cardinals are classified as hard. An automated system might first identify whether a sentence is classified as 'hard' using machine learning and then highlight features whose counts or value are strongly associated with hard sentences to suggest areas for improvement to writers.
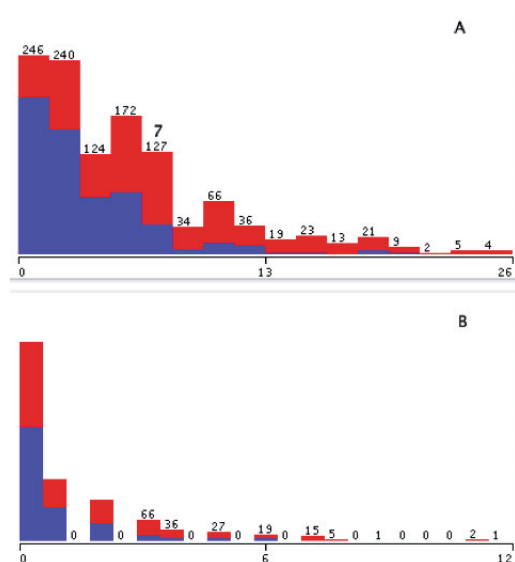


Figure 2: Distribution of easy hard classification by occurrence of feature. A. No. of prepositions occurring in a sentence. B. No. of cardinals occurring in a sentence.

As mentioned above, the underlying algorithm used to select features is based on extracting those most highly correlated with the class output, while not correlated with each other.

We investigated correlations further using the R statistical package.(R-Core-Team et al., 2012) Table 5 shows the features most highly correlated with the language difficulty score used to develop easy/hard classifications. Correlations are shown for the corpora individually and grouped into legal and non-legal categories. Figure 3 also illustrates this data, highlighting the comparative ranking of various predictors of readability for legal vs. non-legal sentences. A number of insights can be drawn.

(1) The highest correlation for the legal corpus is considerably lower of the highest correlation for the non-legal corpus and to a lesser extent for the Brown corpus. Identifying the difficulty of legal language is harder than for non-legal language.

(2) There is a different ordering of correlation of features for the different corpora. The correlation of factors with difficulty of legal language are not the same as for the factors which correlate with language difficulty for general language. For example prepositional phrases and type-to-token ratios are more strongly associated with language difficulty in legal than non-legal sentences relative to other markers. Raw length in words is a relatively weaker marker for legal sentences.
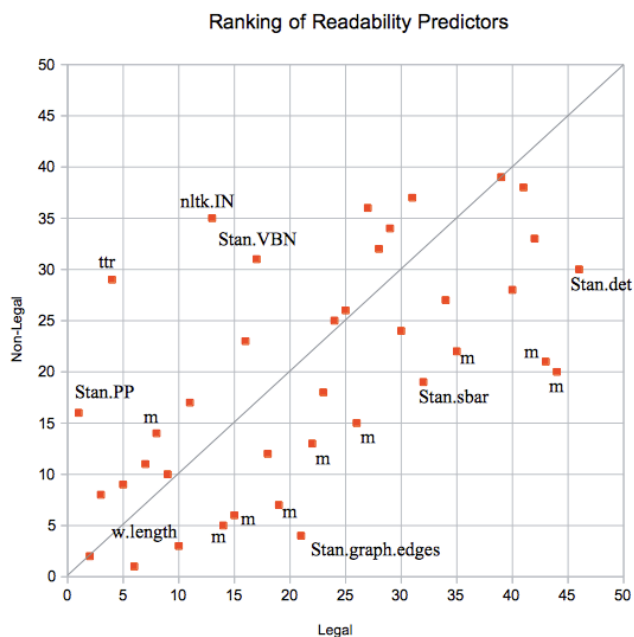


Figure 3: Ranking of predictors of language difficulty for legal and non-legal sentences. 'm' stands for readability metrics, which are generally worse predictors of difficulty of legal sentences than non-legal sentences. Type to token ('ttr') and prepositional phrases (PP) are good predictors for difficulty of legal language but lower ranked for non-legal sentences. Other labelled features are as for Table 5.

(3) Readability metrics have a lower correlation with language difficulty in legal as opposed to general language. This stands to reason as readability metrics were not designed specifically for legal language. In Figure 3, all metrics, except for the SMOG metric have a lower ranking as predictors of difficulty of legal sentences. The result is in line with views sometimes expressed that readability metrics are unsuited to legal language.(Melham, 1993)

(4) Simple sentence length (in line with the results for readability metrics which often use sentence length in their calculation) is a less correlated feature for legal as opposed to non-legal sentences. Average word length also had a low correlation with the whole dataset, although correlation did not attain significance for individual corpora.

(5) Significance tests on correlations point to the need for more labelled data for machine learning. Of all the features used for machine learning, correlations for less than 50 were significant in the legal case. Natural language features have a zipfian distribution. A few features occur very frequently while most features occur very rarely.(Curtotti and McCreath, 2011) A larger dataset is required to attain significance on a larger number of features. This would improve accuracy of machine learning and provide greater insights into the nature of language difficulty.

**Table 2:** <u>**CORRELATION OF NATURAL LANGUAGE FEATURES WITH READABILITY SCORE.**</u> Correlation scores with exclamation mark next to them are not statistically significant. All other correlations have p < 0.05. The lists are ordered by most highly correlated in the legal category. Notes: Stan = stanford parser; dep = dependency; Penn = penn tree bank tagset; degreehist refers to the number nodes in a dependency parsed sentence having a particular degree (i.e. connections); nltk = natural language toolkit. nltk features are normed by sentence length, stan features are raw counts.For ease of reference: PP = prepositional phrase, IN = preposition or subordinating conjunction, NP = noun phrase, STP = full stop, NN = singular or mass noun, JJ = adjective, VBN = past participle verb; POS = possessive ending; DT = determiner, NNS = plural noun, SBAR = subordinating conjunction, RB = adverb, VP = verb phrase, ADJP = adjectival phrase, MD = modal verb, WHNP = 'wh' noun phrase (e.g. where, who, when etc)

| Feature | Brown | Graded | US Code | C.Fed.Reg. | Legal | Non-legal | All |
|---|---|---|---|---|---|---|---|
| StanPhrases.PP | 0.43 | 0.45 | 0.42 | 0.42 | 0.42 | 0.51 | 0.49 |
| StanDep.deghist.2 | 0.45 | NA | 0.39 | 0.40 | 0.40 | 0.58 | 0.49 |
| StanPennTag.IN | 0.47 | 0.51 | 0.43 | 0.40 | 0.40 | 0.55 | 0.50 |
| typetotokenratio | -0.32 | -0.33! | -0.42 | -0.40 | -0.40 | -0.38 | -0.51 |
| StanPhrases.NP | 0.46 | 0.54 | 0.40 | 0.39 | 0.39 | 0.55 | 0.48 |
| StanPenn.subtreecnt | 0.47 | 0.61 | 0.46 | 0.39 | 0.39 | 0.59 | 0.51 |
| nltkPenn.STP | -0.42 | -0.51 | 0.04! | -0.37 | -0.37 | -0.54 | -0.49 |
| rmetric.smog | 0.45 | 0.27! | 0.36 | 0.37 | 0.37 | 0.52 | 0.47 |
| StanPenn.treeheight | 0.48 | 0.48 | 0.38 | 0.36 | 0.36 | 0.55 | 0.48 |
| surface.lengthinwords | 0.46 | 0.61 | 0.35 | 0.32 | 0.32 | 0.58 | 0.42 |
| StanPennTag.NN | 0.38! | 0.33! | 0.40 | 0.32 | 0.32 | 0.47 | 0.46 |
| StanDep.deghist.4 | 0.19 | 0.05! | 0.32 | 0.32 | 0.32 | 0.26! | 0.39 |
| nltkPenn.IN | 0.30 | 0.32! | 0.37 | 0.31 | 0.31 | 0.34 | 0.36 |
| rmetric.fleshkgradlvl | 0.41! | 0.48 | 0.31 | 0.31 | 0.31 | 0.56 | 0.39 |
| rmetric.gunningfog | 0.46 | 0.40 | 0.34 | 0.30 | 0.30 | 0.56 | 0.41 |
| StanPennTag.JJ | 0.32 | 0.30! | 0.32 | 0.30 | 0.30 | 0.42 | 0.39 |
| StanPennTag.VBN | 0.29 | 0.30! | 0.34 | 0.30 | 0.30 | 0.36 | 0.40 |
| StanDep.cliqueav | 0.43 | 0.47 | 0.31 | 0.29 | 0.29 | 0.54 | 0.51 |
| rmetric.ari | 0.41! | 0.45 | 0.31 | 0.29 | 0.29 | 0.56 | 0.38 |
| nltkPenn.POS | -0.07 | 0.01! | -0.17! | -0.29 | -0.29 | -0.05! | -0.08 |
| StanDep.grph.edges | 0.44 | 0.59 | 0.31 | 0.29 | 0.29 | 0.57 | 0.38 |
| rmetric.rix | 0.42 | 0.48 | 0.28 | 0.28 | 0.28 | 0.54 | 0.34 |
| StanPennTag.DT | 0.40 | 0.33! | 0.34 | 0.28 | 0.28 | 0.46 | 0.42 |
| StanDepend.root | 0.31 | 0.31! | 0.26 | 0.27 | 0.27 | 0.39 | 0.33 |
| StanDep.deghist.3 | 0.26! | 0.57 | 0.30 | 0.27 | 0.27 | 0.39 | 0.38 |
| rmetric.lix | 0.32! | 0.43 | 0.32 | 0.26 | 0.26 | 0.52 | 0.41 |
| StanDep.deghist.5 | 0.28! | 0.27! | 0.31 | 0.26 | 0.26 | 0.34 | 0.35 |
| StanDep.avdepdist | 0.26! | 0.30! | 0.15! | 0.25 | 0.25 | 0.36 | 0.31 |
| StanDep.prep_of | 0.22! | 0.28! | 0.17! | 0.25 | 0.25 | 0.35 | 0.31 |
| StanPhrases.S | 0.40! | 0.36! | 0.31 | 0.25 | 0.25 | 0.40 | 0.34 |
| StanPennTag.NNS | 0.28! | 0.21! | 0.19! | 0.24 | 0.24 | 0.33 | 0.30 |
| StanPhrases.SBAR | 0.39! | 0.36! | 0.35 | 0.24 | 0.24 | 0.45 | 0.38 |
| StanPennTag.RB | 0.21! | 0.27! | 0.22 | 0.23 | 0.23 | 0.28! | 0.22 |
| StanPhrases.VP | 0.33! | 0.36! | 0.33 | 0.23 | 0.23 | 0.39 | 0.38 |
| rmetric.flesh | -0.26! | -0.31! | -0.23 | -0.23! | -0.23 | -0.42 | -0.37 |
| StanDep.prep_in | 0.19! | 0.21! | 0.08! | 0.22! | 0.22 | 0.24! | 0.19 |
| StanPhrases.ADJP | 0.27! | 0.05! | 0.18! | 0.22! | 0.22 | 0.18! | 0.22 |
| nltkPenn.MD | -0.01! | 0.09! | -0.16! | -0.21! | -0.21 | 0.09! | -0.04 |
| StanDepend.nn | 0.28! | 0.17! | 0.06! | 0.20! | 0.20 | 0.32 | 0.28 |
| StanDep.grph.avdeg | 0.41! | 0.31! | 0.11! | 0.20! | 0.20 | 0.39 | 0.35 |
| StanPennTag.COMMA | 0.21! | 0.37! | 0.17! | 0.18! | 0.18 | 0.33 | 0.25 |
| StanPhrases.WHNP | 0.31! | 0.28! | 0.23 | 0.17! | 0.17 | 0.36 | 0.26 |
| rmetric.colemanliau | 0.25! | 0.29! | -0.19! | 0.16! | 0.16 | 0.43 | 0.30 |
| rmetric.dalechall | 0.33! | 0.18! | 0.36 | 0.14! | 0.14 | 0.44 | 0.48 |
| surface.avwordlength | 0.04! | 0.02! | 0.04! | 0.13! | 0.13! | 0.19! | 0.23 |
| StanDepend.det | 0.31! | 0.25! | -0.05! | 0.12! | 0.12! | 0.37 | 0.20 |

It needs to be noted that correlation does not necessarily imply causation. Features we have extracted may be markers for other features which might more reasonably be thought to 'cause' greater language difficulty. For example, a correlation between colons and language difficulty does not imply that colons are difficult for people to understand. Rather, we can reasonably infer that the common practice of sub-paragraphing in legal language is associated with greater language difficulty.

The visualization and correlation insights combined, also enable us to suggest probable heuristics for improving the writing of legal language. Visualization tells us whether a sentence with say a certain number of prepositions is likely to be classified as hard. Correlation provides guidance on features which are most correlated with language difficulty (noting that it will be a combination of such features that will make a sentence 'hard'). Heuristics, based on such data studies, could be incorporated into grammar checking software without undue difficulty in some cases, though would require natural language processing in others. Heuristics could also be adjusted to desired ease of reading level. e.g. At least 50% or at least 75% of sentences should be easy. Heuristics could be combined with inline highlighting and suggestions to writers of features that need to be addressed. Some features are suitable for inclusion in written guidelines, if a predetermined standard (such as a majority of easy sentences) is used. The following is an indicative list of heuristics for the writing of regulatory language. These heuristics are drawn from non-ngram features.

(1) Avoid sentences with more than five prepositions. (A standard of six prepositional phrases is already a feature of plain language guidance.(Tanner, 2002))

(2) Keep sentence length below 30 words. (A standard systematically breached in regulatory writing. Short sentences are often recommended in plain language guidance.[6])

(3) Avoid more than 3 cross references to other sections.

(4) Ensure the sentence has enough lexical diversity (different words) to properly explain its concepts.

(5) Avoid more than two conjunctions (and/or). This feature is likely to be a marker for sub-paragraphing.

(6) Avoid more than six determiners (this/that/those/these).

(7) Avoid more than two negatives.

(8) Avoid more than two bracket sets. This would include both cross references and other brackets.

(9) Avoid sub-paragraphing. The presence of a colon is associated with greater language difficulty. Sub-paragraphing is a widely used tool to make regulatory language easier to read. However readers associate such sentences with reading difficulty.

(10) Avoid more than two modal verbs (must/may/shall) in a sentence.

(11) Avoid personal pronouns. Interestingly these are often thought to make regulatory language easier.

(12) Minimise the use of adverbs and adjectives. Note this result requires further investigation. It may be symptomatic of other issues e.g. high use of defined terms that are unfamiliar to readers.

(13) Avoid using the word 'to' more than twice. The word 'to' may be used as a preposition (e.g 'to the north' or as an infinitive verb ('to run')).

(14) Avoid high verb usage.

(15) Aim for at least 50% non-functional words in a sentence (i.e. nouns, verbs, adverbs, adjectives).

It is notable that some of these heuristics are overlapping (e.g. the use of the word 'to' in a prepositional context). Also, some of the correlations may be representative of underlying causes (not actually tested as a feature), but for which the particular feature is a marker. These findings can be compared with plain language guidelines such as those mentioned above. Guidelines specifically drawn up in the context of US regulatory language are the Guidelines for Drafting and Editing Court Rules.[7] The guidelines for example recommend a maximum sentence length of 30 words (in line with our findings). Use of simple words is suggested. However our study finds little correlation between word length and language difficulty in the case of legal language. This could have a number of interpretations - e.g. the words used in legislative sentences are uniformly bad, or they are uniformly good. Singular nouns, in our results, are more associated with language difficulty than plural nouns. This result is contrary to the guidelines. In line with our findings, the guidelines encourage less use of prepositional phrases. Negatives are not mentioned in the guidelines.

# 6.  CONCLUSIONS AND FUTURE WORK

The creation of a corpus of legal language labelled for reading difficulty, has, as far as we are aware, never been undertaken previously. This should provide a useful resource for researchers interested in exploring the readability of legal language further. We show that the accuracy of machine learning results on the task of predicting reading difficulty class on legal language is particularly improved by addition of context free grammar phrases and tagging, where raw counts of these features are used as input to learning. Ngram features are also shown to increase machine learning accuracy. Dependency graph characteristics which we also investigated in this study were not found to be particularly valuable in the task of predicting readability in a machine learning framework. Like other researchers, who note the complexity of readability assessment at sentence level, as compared with document level, generally the more features used the better. This is true for legal sentences as well as non-legal sentences. We were also able to draw insights from machine learning and the associated study of correlations. Among these insights, are empirical confirmation that different factors contribute to the difficulty of legal language, and

---

[6]US Federal Plain Language Guidelines March 2011 http://www.plainlanguage.gov/howto/guidelines/ FederalPLGuidelines/FederalPLGuidelines.pdf

[7]Guidelines for Drafting and Editing Court Rules http://www.uscourts.gov/uscourts/RulesAndPolicies/ rules/guide.pdf

that readability metrics developed for general language are not well suited to legal language. We explore the development of heuristics for readability of legal texts and describe a feasible tool that could be used to assist writers, specifically designed with legal language in mind. Also, in some cases, heuristics can be converted into written recommendations to writers of legal language to improve readability. We provide an indicative list of such heuristics.

There are a number of limitations to the work reported above which are potentially fruitful avenues for future research. It is clear that the collection of user data for tagging the difficulty of legal sentences is feasible and valuable. Increasing the amount of labelled data can be anticipated to further improve machine learning accuracy. Also, it will increase confidence in heuristics, such as the indicative list offered above. The labelling of legal sentences for language difficulty also needs to be extended to other jurisdictions; as different legislative styles may impact differentially on readability and lead to different factors contributing more or less strongly to readability. Ideally language might be collected sufficient to cover a diversity of styles in different jurisdictions. This is not to mention the interesting insights that might be drawn from cross-language studies, such as in the European Union or United Nations contexts, where large parallel collections of multilingual legal texts exist. Finally, the actual development of tools based on the above suggested heuristics, together with the validation of results of their application through user studies, would enable practical improvement of legal language over time. Such a tool could, for example be applied to guide a rewriting of the most read portions of the US code, given that our previous research has identified that such focussed revision provides disproportionate benefits. An entire rewrite is unnecessary. The heuristics developed above are derived largely from subjective evaluations by users of sentences. It remains an open question whether these subjective evaluations are highly correlated with objective comprehension of legislative sentences. Our cloze data suggest that the correlation is moderate (i.e. around 50%). While previous research suggests a high correlation between cloze tests and comprehension tests, whether this holds true in respect of legal language has not been subjected to specific investigation, as far as we are aware. The application of legal informatics to the improvement of legal language continues to provide an open research field where valuable work can be undertaken.

## References

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

Jin Cong and Haitao Liu. Approaching human language with complex networks. *Physics of life reviews*, 2014.

M. Curtotti and E. McCreath. A corpus of australian contract language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law 2011*, 2011.

Michael Curtotti and Eric McCreath. A right to access implies a right to know: An open online platform for research on the readability of law. *Journal of Open Access to Law*, 1(1), 2013.

Michael Curtotti, Wayne Weibel, Eric McCreath, Nicolas Ceynowa, Sara Frug, and Tom Bruce. Citizen science for citizen access to law. *Journal of Open Access to Law*, 3 (1), 2015.

Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. Assessing the readability of sentences: Which corpora and features? *ACL 2014*, page 163, 2014.

Johan Falkenjack and Arne Jönsson. Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 114–122, 2014.

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, 1964. URL `http://www.hit.uib.no/icame/brown/bcm.html`.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software. *SIGKDD Explorations*, 11 (1), 2009.

Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

HaiTao Liu and Jin Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.

HaiTao Liu and WenWen Li. Language clusters based on linguistic complex networks. *Chinese Science Bulletin*, 55 (30):3458–3465, 2010.

D. Melham. Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs, 1993.

Joakim Nivre. Dependency grammar and dependency parsing. *MSI report*, pages 1–32, 2005.

R-Core-Team et al. R: A language and environment for statistical computing. 2012.

Edwin Tanner. Seventeen years on: Is Victorian legislation less grammatically complicated. *Monash UL Rev.*, 28:403, 2002.

# Citizen Science for Citizen Access to Law

Michael Curtotti[*], Wayne Weibel[+], Eric McCreath[*], Nicolas
Ceynowa[+], Sara Frug[+],Tom Bruce[+]
[*]*Research School of Computer Science, Australian National University*
[+]*Legal Information Institute, Cornell University Law School*

**Abstract.**

Over 2014, the Cornell University Legal Information Institute and the Australian National University worked with users of the Cornell LII site in a citizen science project to collect over 43,000 crowdsourced assessments of the readability of legal and other sentences. Readers ("citizen scientists") on legislative pages of the LII site were asked to rate passages from the United States Code and the Code of Federal Regulations and other texts for readability and other characteristics. They were also asked to provide information about themselves as part of the audience that uses legislation online. The overall aim of the project was to develop empirical insights into characteristics of law that may make it easy or hard to read for the audience that use it. Also, the project aimed to assess machine learning for automatically predicting readability of legal sentences at sentence level.

A major focus of this paper is to report results and insights from demographic data collected during the study. Understanding the audience which reads the law is directly relevant to readability - as the relevant question is readable by whom? Who are the citizens for whom "citizen access" might be enhanced? The paper also describes methods used to rank sentences by readability, using the data provided by citizen scientists. Finally, the paper reports initial tests on the viability of machine learning as a means of predicting readability in advance. The exploratory machine learning results reported here will be extended in further work reported in a future paper.

The research provides insight into who uses legal rules and how they do so. We draw conclusions as to the current readability of law, as well as the spread of readability among legal rules. The research creates a dataset of legal rules labelled for readability by human judges. As far as we are aware, this research project is the largest ever study of readability of regulatory language and the first research which has applied crowdsourcing to such an investigation.

**Keywords:** readability, legislation, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, cloze testing, crowdsourcing, citizen science

## Table of Contents

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

Citizen Science for Citizen Access

## 1. Introduction

*Citizens should be able to know and understand the law that affects them. It is unfair to require them to obey it otherwise.New Zealand Law Reform Commission & Office of Parliamentary Council*(NZ, 2007)

The readability and usability of law has long attracted critical attention from users, providers, researchers and others. This paper reports research which seeks to strengthen the empirical foundations for assessing the reading difficulty of legal rules with the ultimate aim of enhancing "citizen access" to law.

In 2013 the UK Parliamentary Counsel observed:

*Legislation affects us all. And increasingly, legislation is being searched for, read and used by a broad range of people. It is no longer confined to professional libraries; websites like legislation.gov.uk have made it accessible to everyone. So the digital age has made it easier for people to find the law of the land; but once they have found it, they may be baffled. The law is regarded by its users as intricate and intimidating.*(OPC-UK, 2013)

In 1992 it could be said that only 'a lunatic fringe' in the public would read legislation.(Krongold, 1992) Whether or not true then, by 2013, the UK Parliamentary Counsel could confidently state that it was no longer necessarily the case that readers of legislation were legally qualified. They report an audience of two million unique visitors per month for the legislation.gov.uk site.(OPC-UK, 2013)

Most of this paper discusses a project which applies "citizen science" to the problem of making law more readable. Two sub-problems in particular are addressed, building on the crowdsourced data collected for this research project. What are the characteristics of the audience which reads the law? Which parts of legal language are difficult for its readers? Both these sub-problems are empirical in nature. Much work - including empirical work - has been done in the past with legal language (for example in the plain language movement). The use of crowdsourced techniques in a citizen science project has not been applied to this task, as far as we are aware.

While amateur science has a long and respectable history (for example in the field of astronomy), the recency of the phrase "citizen science" is underlined by its addition to the Oxford English Dictionary only in June 2014. The Dictionary defines it as *"scientific work undertaken by members of the general public, often in collaboration with or under the*

*direction of professional scientists and scientific institutions".*[1] Other definitions have been proposed, and one that approximates our own project in part is the following *"the participation of nonscientists in the process of gathering data according to specific scientific protocols and in the process of using and interpreting that data."* (Lewenstein, 2004; Wiggins and Crowston, 2011)

Our citizen science project uses "crowdsourcing", another recently invented term (Jeff Howe in 1996). The term (although definitionally contested) expresses the idea of engaging a large number of people outside an organisation to undertake a task or solve a problem, typically online (i.e. using web technologies). Like citizen science, precursors to crowdsourcing can be found well before the 21st century. The arrival of the internet has greatly amplified the opportunity for individuals and organisations to work together towards a shared goal and many crowdsourced projects are well-known. Crowdsourcing via the web has been applied in many fields, including in citizen science projects: for example classifying galaxies, folding proteins and identifying cometary dust collected in outer space. (Howe, 2006; Brabham, 2008; Doan et al., 2011; Hand, 2010; Asmolov, 2014; Poblet et al., 2014)

In the case of our study, citizen science has been used, not only to study the language of the law, but also to learn more about people who use that language, as well as their experience of that language. The research thus engages citizen scientists in research which involves learning more about themselves as well as objective characteristics the 'data out there'. This is necessary in the context of the goals involved, as any exercise in enhancing readability is only meaningful if it addresses readability in the context of the experience and needs of the audience for given written materials.

To undertake our study we prepared a corpus of around 1250 randomly selected sentences from four different collections of English language:

**(a)** 139 sentences drawn from graded reading materials;

**(b)** 112 sentences drawn from the Brown corpus of English;

**(c)** 500 sentences from the United States Code; and

**(d)** 500 sentences from the US Code of Federal Regulations.

The Brown corpus is a balanced collection of written American English and is used as a reference point for 'normal American En-

---

[1] Oxford English Dictionary http://public.oed.com/the-oed-today/recent-updates-to-the-oed/june-2014-update/new-words-notes-june-2014/ and http://www.oed.com/view/Entry/33513

Citizen Science for Citizen Access

glish'.(Francis and Kucera, 1964) The graded reading material is drawn from 'readers' for language learners.[2] This corpus represents a modified written English simplified to be accessible to readers with different levels of reading skill. Both the Brown and Graded corpora provide reference points for calibrating and validating assessments of the legislative corpora. The US Code and the Code of Federal Regulations constitute the primary subjects of study. It may be noted that legislative rules (such as those drawn from the US Code and Code of Federal Regulations) have something in common with the graded corpus. They are also a form of modified English. Although simplicity is not the primary goal of legislative drafting - clarity, simplicity and readability are subsidiary goals that the creators of legislative texts pursue and regard as important.(Bowers, 1980; of Victoria, 1990; Melham, 1993; Tanner, 2002; OPC-Australia, 2003)

To obtain human judgements about the readability and other characteristics of the test sentences described above, we created an online interface which invited readers at the LII Cornell website to become research participants in a citizen science project. Participants were asked to provide objective and subjective responses to the test sentences. They were also asked to provide broad demographic information about themselves. Participants were, in particular, visitors who had browsed to a section, regulation or rule page of the US primary or secondary legislation at the LII Cornell site. The research participants are therefore the readers of legislative rules within the US context (i.e. the audience for whom readability of online legislative material is relevant).

Each participant was presented with a test sentence and they were asked to provide one of three alternative assessments of the test sentence.

(a) The participant might be asked to complete a likert question asking how strongly the participant agreed or disagreed with a statement as to how easy or hard the sentence was to read.

(b) Alternatively the participant would be presented with a cloze deletion test which asked the participant to guess up to ten missing words in the sentence.

(c) Otherwise, the participant was asked to complete a semantic differential test which asked the participant to rate the sentence on seven point scale against ten pairs of semantic opposites such as "readable-unreadable", "usable-unusable", "attractive-repulsive".

---

[2] Graded reader sentences extracted from graded reader passages downloaded from http://www.lextutor.ca/graded/. No longer available at time of publication. A copy of the corpus can be obtained for research purposes by contacting the authors

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

If they wished to do so, participants could assess multiple sentences, until opting out of the study. Also participants were provided with the option of providing demographic data. This included information about their gender, age, linguistic background, place of birth, educational attainment and professional background.

In addition to the foregoing, Google Analytics data on usage of LII legislation pages was also collected and analysed.

Each sentence was rated for its "language difficulty" by combining user ratings using principal components analysis and other methods. Principal components analysis is a mathematically robust method for combining many variables about an instance of data into a smaller number of variables.

This made it possible to order the sentences by language difficulty and assign them to "easy" or "hard" classifications for later use in machine learning. Natural language characteristics (such as sentence length, parts of speech and type to token ratios) were extracted from the test sentences themselves. These features were used in preliminary machine learning tests to examine how accurate machine learning would be in predicting the assigned classes.

Some of our key results are described below. For people who read legislation online, our results included the following.

**(a)** On the LII Cornell site, a very small proportion of the US Code is read very often, while the bulk of the Code is read very rarely.

**(b)** Among our research participants, legal professionals (including law students) were a minority.

**(c)** In proportional terms women, those without tertiary education and Spanish speakers are under-represented among those who participated in the study.

**(d)** The law was easier to read for legal professionals and law students than for other others who participated in the research.

For readability, our results include the following.

**(a)** The project demonstrates the feasibility of long-term collection of online assessments of the readability of legal texts.

**(b)** From user assessments provided, we were able to rank approximately 1000 legislative sentences by language difficulty.

**(c)** In initial application of machine learning algorithms overall accuracy (while not very high) exceeded accuracy of traditional readability metrics.

Citizen Science for Citizen Access

We draw a number of conclusions from our results. It is already known that the *direct* audience of legislative materials now extends far beyond lawyers and the legally trained. The results of our study are interesting in providing a quantitative indication of the modern online audience for legislation. That the non-legally trained were the majority of respondents in our study is significant. It provides quantitative validation that non-lawyers are a *substantial* audience for legislative materials. It suggests that they may now be a majority among readers of such materials. As this result may have other explanations, further studies will be required (including on other sites) to determine whether this is in fact the case. The under-representation of women among research participants is also interesting. Again it may have a variety of explanations and merits further study. The under-representation of those without tertiary education and spanish speakers is a result that might be expected, but in this case points to the relevance of asking questions about citizen access, as a likely reason is that the under-representation is a marker for lack of access.

The result that the law is easier for lawyers than non-lawyers is not surprising. Traditionally, the law has been written by lawyers, for lawyers. It is interesting however to be able to quantify the difference. In cloze deletion tests, the legally trained outperformed the non-legally trained on legal, but not non-legal, sentences. The difference was significant, but the effect size was small. Using traditional cloze deletion test analysis, the results suggest legal language is hard for all audiences (including the legally trained). For members of the public the difficulty level was 'frustrational'.

It is interesting to note the wide spread of readability in legal sentences. This suggests that there is no inherent reason why legislative sentences must be difficult. Many legislative sentences are not. For machine learning, our results confirm for the legislative field that readability metrics can readily be improved on. Results are nonetheless preliminary and we intend to extend analysis in a future paper. We leave further discussion of results to the conclusions.

Section 2 discusses related research and theoretical frameworks. Section 3 provides an overview of the study and how it was carried out. Section 4 discusses demographic data. Section 5 discusses the methods used to rank and classify sentences for language difficulty. Section 6 discusses results obtained from initial exploratory application of machine learning.

## 2. Related Work

The subsections which follow provide a background to our research. Given the multidisciplinary nature of our work, a number of fields from law, research methods, statistics and computer science are relevant. The fields we address are access to law; readability; plain language; readability applied to legislation; citizen science; crowdsourced research on readability; natural language processing; machine learning; assessing the reading difficulty of sentences; likert testing; cloze testing and semantic differentials. Necessarily the coverage of any particular area is as brief as possible. Nonetheless, the aggregate discourse is quite long and readers who are already familiar with these fields may wish to skip all or part of this discussion and go to Section 3 and following which describes our study and results.

### 2.1. Access to Law

Access to law has a number of possible meanings. The New Zealand Law Commission and the New Zealand Parliamentary Counsel's Office identify three.[3] Firstly, access in the sense of 'availability' to the public (such as via hard copy or electronic access). Second, 'navigability' - the ability to know of and reach the relevant legal principle. Finally, 'understandability' - that *'the law, once found, [is] understandable to the user.'* (NZ, 2008) We are primarily concerned with access to law in this third sense.

In 1983, a Parliamentary draftsman, F.A.R. Bennion, observed: *"It is strange that free societies should ... arrive at a situation where their members are governed from cradle to grave by texts they cannot comprehend."* The startling character of this observation arises from an incongruity of notions of 'freedom' and 'democracy', with the reality that most members of society are unable to access the meaning of laws which set out their rights and responsibilities as citizens. Ironically, Bennion himself believed that laws were written for lawyers and legal professionals and nothing could really be done about it.(Curtotti and McCreath, 2012)

This is not a view that is widely held and a number of sound democratic and other reasons have been advanced as to why laws should be understandable by all those to whom they are addressed.

---

[3] While our study was conducted on an American legal website using American legal text, the case for greater readability of legal materials is general across the english speaking world, and indeed beyond. Accordingly, our discussion draws on the most helpful materials. wherever we have found them.

Citizen Science for Citizen Access

**Arguments from Rule of Law:** One argument is based on the rule of law. If laws cannot be understood, it becomes difficult to sustain the rule of law, as the laws themselves are inaccessible. Implicit in this rationale is that the rule of law is in itself a social good: a social good which is frustrated by poor communication.

**Arguments from Equity:** Another argument is based on fairness: that to expect citizens to obey rules they cannot understand is unfair.

**Arguments from Legislative Effectiveness:** From the viewpoint of the legislator, adopting laws which cannot be understood is inefficient, at best, or futile, at worst. The legislator presumably wishes to communicate so as to optimally achieve its intent.

**Arguments from Economic Efficiency:** From the viewpoint of economic efficiency, the language should result in minimal regulatory burden. Efforts at tax law simplification are of this kind. Beyond preserving resources for other uses, implicit in this kind of reasoning is that freedom is a social good - limitations of which should only be imposed to the extent necessary to achieve a regulatory intent.

**Arguments from Audience:** As the Good Law initiative notes, the audience of legal rules has changed. Laws are available on the web and they are read by everyone. Laws should be written for the audience which reads it. Implicit in this rationale is a customer or citizen service orientation. Law is a service provided to its end 'users' and should be optimally *designed* to meet the needs of its users.[4]

**Arguments from the Commons:** A 'commons' argument regards the law as a form of property which in a sense 'belongs' to everyone. This principle underlies the founding documents of the Free Access to Law Movement. The Declaration on Free Access to Law states: *"Public legal information ... is part of the common heritage of humanity ... [it] is digital common property and should be accessible to all on a non-profit basis and free of charge."*[5]

**Arguments from Rights:** Close to the commons argument are rights arguments. Some authors argue that there is, or should be a 'right to access the law'.

---

[4] Note that the demographic results that we describe below provide an empirical description of the user base of the US legislative material.

[5] http://www.worldlii.org/worldlii/declaration/

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

**Arguments from Democracy:** As implicit in Bennion's observation cited above, open access to law can also be argued from a democratic viewpoint.[6]

Of course, these arguments apply to access to law in all three of its senses. For example, the Free Access to Law Movement began with a focus on access to law in the sense of universal free online availability.

Those who create the law are well aware of the need for it to be as accessible as possible. The Australian Office of Parliamentary Counsel put it this way in its plain language drafting guidance.

> *We also have a very important duty to do what we can to make laws easy to understand. If laws are hard to understand, they lead to administrative and legal costs, contempt of the law and criticism of our Office.* (OPC-Australia, 2003)

## 2.2. What is readability and how is it measured

DuBay reviews a number of the definitions that are offered for readability: *'readability is what makes some texts easier to understand than others'*; *'the ease of understanding or comprehension due to the style of writing'*; *'ease of reading words and sentences' as an element of clarity*; *'the degree to which a given class of people find certain reading matter compelling and comprehensible'*; and *'The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it.*(DuBay, 2004)

From the early 20th century researchers of language began to develop ways to measure the readability of language. A variety of "readability metrics" were developed. Such measures were used by educators to rank material for appropriate age levels. Writers also used the metrics to make their writings more usable for their intended audiences.(DuBay, 2004)

Reading measures such as the Flesch, Flesch-Kincaid, Gunning, Dale-Chall, Coleman-Liau and Gary-Leary are among the more than 200 formulas which have been developed to measure the readability of text. These formulas (although varying in formulation) address two underlying predictors of reading difficulty: semantic content (i.e. the vocabulary) and syntactic structure. Vocabulary frequency lists and sentence length studies both made early contributions to the developments of formulas. The Flesch formula calculates a score for reading

---

[6] For a more detailed description of the principles of access to law discussed above see: (Curtotti and McCreath, 2013).

Citizen Science for Citizen Access

difficulty using average sentence length and average number of syllables per word. Formulas of this kind are justified on the basis of their correlation with reading test results. For example, the Flesch formula correlated at levels of 0.7 and 0.64 in different studies carried out in 1925 and 1950 with standardised graded texts.(DuBay, 2004)

Most work undertaken on readability assesses passages of a given length (often 100 or more words). This arose because most of the creators readability metrics were seeking to use them to rate passages for inclusion in educational materials. The approach is ill suited to identifying specific linguistic features that contribute to difficulty of legal language. In a larger passage, the metric is spread over a broader vocabulary - and over a potentially large number of syntactic constructs. Greater resolution is required to be able to distinguish specific language elements contributing to language difficulty.

The uses and abuses of readability formulas have been widely debated. Readability metrics were not conceived as ways of improving the writing of text, rather they were designed to help teachers select appropriate existing texts for children of different ages.(Woods et al., 1998)

In 1993, a report to the Australian Parliament (having reviewed use of readability metrics) expressed a lack of confidence in using readability metrics on legislation. The report commented:

> 'Testing for the readability of legislation by using a computer program is of limited value. The most effective way of testing legislation is to ask people whether they can understand it - a comprehension test.'(Melham, 1993, p xx)

## 2.3. Plain language, readability and legislation

Concerns about the readability of law are far from new. In England, against the resistance of the legal profession, legal language had to be prized from the medieval but firm grip of French, Latin and technical legalese. Again in Georgian times there was a 'clamor for legible [legal] English'. Again the profession opposed reform, in that case with success.[pp 124 et seq, pp133 et seq](Mellinkoff, 1963) In the 19th century, laws of the British Parliament still consisted of great slabs of discursive text. In the early-nineteenth century, Jeremy Bentham (credited with being one of the writers influencing later reforms) vociferously critiqued the problems of legislative drafting. His critique included the failure to use such obvious tools as division of legislative texts into digestible portions and section numbering to aid retrieval.(Bowers, 1980), [pp 250-251](Bentham, 1843) Practices such as section numbering and the breaking up of text were officially endorsed with the passage of Britain's

first Acts Interpretation Act in 1850. These reforms were bedded down after the first parliamentary drafting office was established in the late nineteenth century. Such offices reformed legislative drafting, including by structuring Acts in parts and use of sub-paragraphing. (See for example [p 250](Bentham, 1843),(Evans and Jack, 1984; Renton, 1975; Bowers, 1980))

In modern times, the United States also pursued plain english in the law, building on its own history of concern about legal english. In 1963 David Mellinkoff's book *The Language of the Law* appeared with the aim of "making an existing language better perform its function". In the 1960's and 70's, plain language began to appear in some insurance and consumer contracts. In the 1970's and 80's, state and federal laws began to mandate the use of readily understandable language in legal documents.(Friman, 1994) "In June 1998, President Clinton directed all federal agencies to issue all documents and regulations in plain language."(DuBay, 2004) The Plain Writing Act of 2010 mandates that US government agencies use language the public can understand and an executive order issued by President Obama in 2011 requires regulations to be "accessible, consistent, written in plain language, and easy to understand."[7]

Now it is possible to state that legislative drafting offices frequently commit to plain language as a goal they pursue.(Kimble, 1994; OPC-Australia, 2003)

Proponents of plain language cite extensive empirical studies validating the benefits of plain language. In the research field, extensive work has been undertaken to study the effect of improving legislative language.

An early example was a study reported in 1984 in which cloze testing was undertaken on several samples of legal text including legislative language. One hundred generally highly educated non-lawyers (28% had undertaken some postgraduate training) were tested. The group averaged 39% accuracy, a result close to 'frustrational' level for cloze testing. Ten participants, who had only high school education, experienced even greater difficulty, averaging 15% - a result consistent with total incomprehension.(Benson, 1984)

In 1999, Harrison and McLaren studied the readability of consumer legislation in New Zealand, undertaking user evaluations, including cloze tests. The study found traditional readability metrics to be unreliable. The results of cloze testing extracts from the legislation led to the conclusion that the legislation would require explanation before being comprehended at adult level. For young adults (aged 18-34),

---

[7] Plain Language: It's the Law. http://www.plainlanguage.gov/plLaw/

Citizen Science for Citizen Access

comprehension levels were even lower (within the frustrational level). Participants complained of the length of sentences and most felt there was a need for some legal knowledge to understand the text. All felt the text should be made easier.(Harrison and McLaren, 1999)

In the early 1990's Australia, New Zealand and the United Kingdom pursued tax law simplification initiatives which involved rewriting at least substantial portions of tax legislation. In Australia's case cloze testing on a subset of the work was inconclusive. Participants found both the original language and the rewritten language difficult.(James and Wallschutzky, 1997) Smith et al., reviewing the effectiveness of the same program, concluded that results fell 'far short of an acceptable bench-mark'. They used the Flesch Readability Score, finding that readability of sections of tax law replaced in the tax law improvement program, improved on average from 38.44 to 46.42 - a modest improvement. Even after improvement, the legislation remained difficult to read. Over 60% of the revised legislation remained inaccessible to Australians without a university education.(Smith and Richardson, 1999)

A 2003 review of the Capital Allowances Act in the UK, which was rewritten as part of the UK's tax law improvement program, undertook interviews with a number of professional users. These professionals in general responded that the new legislation was easier to use and more understandable.(OLR, 2003)

A similar review of the Income Tax (Earnings and Pensions) Act, also carried out in the UK, again found that the interviewed group (primarily tax professionals), were largely positive about the benefits of the simplification rewrite. They expressed the view that the revised legislation was easier to use and understand, although also noting the additional costs of re-learning the legislation.(Pettigrew et al., 2006)

A 2010 study of the effects of the tax law simplification in New Zealand used cloze testing to determine whether the simplification attained its goals. They reported that most of their respondents (mainly respondents unfamiliar with the tax system) found the cloze testing either difficult or extremely difficult. They found that the older (un-amended) Act was the least difficult - a finding contrary to their expectation given earlier research in New Zealand. This they attributed to the nature of the selections from the older legislation. The overall average cloze results was 34.17, with unfamiliar respondents achieving 30.86%. They note that less than 25% of their subjects were able to exceed the instructional level of 44%.(Sawyer, 2010)

A study in Canada carried out usability testing on plain language and original versions of the Employment Insurance Act. Members of the general public and expert users were recruited to carry out testing.

All respondents, particularly those from the general public, found navigation and comprehension difficult, irrespective of version. Also, for all versions respondents faced difficulty in understanding the material. These findings indicated that while plain language reduced difficulty it did not eliminate it. Nonetheless participants preferred the plain language version and found it easier to use.(GLPi and Smolenka, 2000)

Tanner carried out empirical examination of samples of Victorian legislation, assessing them in light of plain language recommendations of the Victorian Law Reform Commission made 17 years earlier. In a study of six statutes, he found that the average sentence length was almost double that the Commission recommended (i.e. an average of 25 words). Also, over time, sentence length had increased. Although he also notes some improvements, he concludes: *"The net result is that many of the provisions are likely to be inaccessible to those who should be able to understand them. This is because the provisions 'twist on, phrase within clause within clause'."*(Tanner, 2002)

An empirical study of the usability of employment legislation in South Africa found that respondent accuracy improved considerably with a plain language version of the legislation. The respondents who were drawn from year 11 school students averaged a score of 65.6% when tested on the plain language version, whereas the control group scored an average of 37.7%.(Abrahams, 2003)

The empirical readability research suggests two conclusions. Firstly writing in plain language assists comprehension of legislation. Secondly legislation is generally incomprehensible or difficult to read to large sections of the population, even in those cases where plain language revision has been undertaken.

## 2.4. CITIZEN SCIENCE AND CROWDSOURCING FOR ASSESSING LANGUAGE DIFFICULTY

As noted in the introduction, citizen science is not new. However, the availability of the internet and software has made engaging volunteers in scientific work far easier than it was in the past. Wiggins and Crowston undertake an extensive review of citizen science projects in a number of dimensions. They identify five mutually exclusive types of projects: action, conservation, education, virtual and investigation. Action projects are focussed on engaging volunteers to address local issues. Conservation addresses natural resource management. Investigation refers to scientific investigation in a physical setting. Virtual projects have similar goals to investigation projects, but in an online setting. Education is primarily concerned with education and outreach. They also note that citizen scientists may be engaged in data collection

Citizen Science for Citizen Access

and analysis, participation in project design and in drawing conclusions and disseminating results. Citizen science projects are typically organised in a top down fashion by a scientific team and volunteers are recruited to assist in the conduct of the project. This is also true virtual projects which typically have a 'top down' organisation. However, sometimes citizen science are organisationally 'bottom up', though this is largely limited to local projects. Scientific issues arise for all types of citizen science. For virtual projects the primary scientific challenge is scientific validity of results and achieving a design that maintains participant interest. Success depends on reaching a critical mass of contributors. The primary approach to ensure validity is replication of results. (Wiggins and Crowston, 2011)

Citizen science projects (particularly those carried out online) can be appropriately considered a form of crowdsourcing. Crowdsourced typologies are similar to those for citizen science projects, particularly as to how the crowd is involved. Poblet et al identify a hierarchy of crowd involvement, based on the type of data that is being crowdsourced. At the base, the crowd may merely serve as sensors (as in data automatically generated by mobile devices). The crowd may be "social computer" - i.e. generators of data later available for assessment (as an indirect rather than intended outcome). The crowd may serve as reporters (i.e. information generators). The crowd may be microtaskers (i.e. performing specific tasks over raw data).(Poblet et al., 2014) Asmolov discusses a broader typology of crowdsourcing, extending the analysis to the question of the level of crowd engagement. At one end of the spectrum is full organizational control - at the other the organization is merely incidental to the crowdsourced activity. The character of crowdsourcing is also disputed: is it the wisdom of the crowd - or the crass capitalist exploitation of unpaid workers? Is it participatory or is it exploitative? Of course, different projects may have one or other of these characteristics. Key to understanding crowdsourcing is what it does: it enables action through accessing resources of the networked crowd (e.g. intellectual, computational, physical or financial).(Asmolov, 2014)

In our own study, all the citizen science dimensions discussed above are in play. Our research is firmly within the virtual space and displays the characteristics mentioned by Wiggins and Crowston for that space. Organizationally, the project was framed in a top down fashion. Citizen scientists were primarily asked to participate in assessing data on a platform designed without their involvement. The platform was designing in a way which was hoped to maintain interest; providing a variety of tests, as well as exploring different ways of assessing readability through crowdsourced evaluations. After data collection, careful

review of data was required to remove confounding inputs (i.e. data validation). Replication of input was a primary means of controlling for 'bad' data. The size of the participant base became an issue, limiting how far the project could go (i.e. how quickly data could be collected). It would be interesting to attempt to expand the scope of citizen scientist participation in future projects, though this may skew participation away from a balanced reflection of the audience for online legislative materials.

The term citizen science has not been used in the readability sphere, nonetheless there are a very small number of projects (under the rubric of crowdsourcing), which also amount to citizen science projects. We have only been able to identify two which were focussed on validating crowdsourcing as a method for readability studies.

De Clercq et al. evaluate the effectiveness of crowdsourcing as a method of assessing readability. They compared the accuracy of crowd-sourced human judgements of the readability to those of expert judges, finding a high level of agreement in readability ranking between the experts and crowdsourced users. Crowdsourced users were presented with two randomly selected texts of one to two hundred words and invited to rank them by readability. Expert teachers, writers and linguists were given a more complex task of assigning a readability score to each presented text. The researchers concluded that crowd sourced user judgements and expert judgements were highly correlated as to readability ranking. They found also that readability metrics had a lower correlation with both these two judgement sets.(De Clercq et al., 2013)

A more general study by Munro et al. concluded that there was a high correlation between traditional laboratory experiments and crowd-sourced based studies of the same linguistic phenomena. Among their conclusions was that crowdsourced judgements closely correlated with cloze testing results. (Munro et al., 2010)

We are unaware of any previous studies which have used crowd-sourcing to assess the readability of legislative texts.

## 2.5. Natural language processing and machine learning

Recent years have seen a growing body of research applying natural language processing and machine learning to assessing the readability of text. The term 'natural language processing' represents the capacity of computers to hold and analyse potentially vast bodies of text. Natural language processing typically transforms natural language into collections of variables representative of the characteristics of the natural language. Such characteristics range from the raw text itself, to

Citizen Science for Citizen Access

representations of the syntactic and vocabulary characteristics of a text. Such characteristics are then available of further processing or analysis.

Machine learning is a well elaborated process. In summary, it seeks to make predictions based on a body of data. Characteristics from that data is extracted as 'input features' and provided to one or more of a variety of machine learning algorithms. The most common goal is for the algorithm to be able, based on patterns in the data, to return a model which predicts the class of a previously unseen item of data. Machine learning includes both 'supervised' and 'unsupervised' learning. In supervised learning, training data (already labelled with the appropriate classifications) is provided to 'train' the learning algorithm. In the unsupervised case, the machine learning algorithm tries to separate the data into natural groupings based on clusterings of features.[8]

Both natural language processing and machine learning have been applied to automatically predict readability. An exhaustive review is not carried out here but a number of aspects of particular interest are highlighted. A key question is what features might assist us in assessing readability? Studies have systematically examined sets of features for their utility in assessing readability. The most easily extracted features are readability metrics and 'surface' features such as average sentence length, average word length, average syllable length, capitalisation and punctuation. Other features studied include lexical features such as vocabulary and type/token ratio,[9] parts of speech frequencies, ratio of content words to function words, distribution of verbs according to mood, syntactic features such as parse tree depths, frequency of subordinate clauses, ngram language models, discourse features, named entity occurrences, semantic relationships between entities and anaphora occurrences.(Si and Callan, 2001; Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Heilman et al., 2008; Pitler and Nenkova, 2008; Kate et al., 2010; Aluisio et al., 2010; Feng et al., 2010; Dell'Orletta et al., 2011; Kauchak et al., 2014) A good overview of the state of the art is provided by Collins-Thompson's survey article on readability research using machine learning.(Collins-Thompson, 2014)

Applying natural language processing and machine learning to predict readability has made considerable progress over the last decade or so. Studies such as those referenced above demonstrate that prediction of readability of text can be improved by incorporating higher level

---

[8] See Bird et al. for an accessible and practical introduction to natural language processing. Chapter six introduces machine learning for classifying text.(Bird et al., 2009)

[9] A 'type' is, say, the word 'red' and a token is any word. So in the phrase "the cat sat on the mat" the type to token ratio is 5/6, as the word 'the' occurs twice.

linguistic features into predictive models. It is also notable that only initial steps have been taken to apply findings in this field to identifying reliable methods to *improve* readability of text.

A limitation of such methods is that without a considerable body of labelled data, it is difficult to attain high levels of accuracy with machine learning. The use of crowdsourced methods enables this problem to be addressed.

## 2.6. Assessing the readability of sentences

Historically, as we have seen above in section 2.2, readability has been addressed at the level of at least a passage of text. Klare notes that readability metrics are designed for larger blocks of text providing a connected discourse. They won't work well on disconnected fragments or single sentences.(Klare, 2000) Fry is one of the few who as early as 1990 sought to create a metric better suited to short passages. At the time, most metrics were designed for use with passages of 300 words or more. Fry particularly noted that such 'short passages' were important in materials such as 'science textbooks, math textbooks, passages used in tests, manufacturers' warranties, and rules and procedures in driver's training booklets.' Fry's new metric could be applied to passages with 40 words or more. However Fry stated that the new metric was only appropriate for passages with at least 3 sentences - making it inapplicable to detecting readability of single sentences.(Fry, 1990)

For this study, sentences were chosen as the unit of study. A rationale for this choice is that the sentence is the basic unit of content for legal rules. That is, in many jurisdictions, each rule is contained in a separate sentence. Another reason for this choice is that longer passages are not sufficiently granular to automatically identify features which contribute to reading difficulty. Without this level of granularity, it is difficult to automate recommendations for improving how materials are written.[10]

Studies exploring language difficulty at sentence (or smaller) level have only emerged recently; with the availability of computational tools which make it more practical.

A number of studies exist which seek to explore smaller units of text. Kanungo and Orr carry out a study of snippets of text returned as web search results which are either sentences or sub-sentences. They present a study involving 5000 human judgements of the readability of such short text fragments. They apply machine learning using a gradient boosted decision tree as the learning model. Their study assesses a number of features (e.g. fraction of capitalisation and fraction

---

[10] It should be noted however that semantic meaning is often connected across sentences. Analysis of this broader level of meaning is lost at sentence level.

Citizen Science for Citizen Access

of search terms) as predictors for the reading difficulty assigned by human judges. They also assess traditional reading metrics such as the Fog, SMOG, Flesch Kincaid metrics. They find that metrics had virtually no correlation with human judgements of the readability of search results . On the other hand, the Pearson correlation R of their boosted decision tree model correlated at around 63%. This study illustrates the inapplicability of traditional metrics to short language segments and to specialised language (i.e. search results in this case).(Kanungo and Orr, 2009)

Dell'Orletta et al note that much work on readability in the natural language processing field is focussed at document level but that such methods are unreliable at sentence level. They study readability at sentence level on the rationale that this would be useful for text simplification. (A rationale that applies in the context of enhancing access to law). They develop a model capable of accurately labelling sentences for reading difficulty with 78% accuracy. Their model includes a range of linguistic features beyond those traditionally used in readability formulas.(Dell'Orletta et al., 2011)

Sjoholm is another researcher who assesses the readability of sentences. He notes the absence of existing metrics for predicting readability at sentence level. He builds on previous studies by developing a probabilistic soft classification approach that rather than classifying a sentence as 'hard' or 'easy' gives a probability measure of membership of either class.(Sjöholm, 2012)

## 2.7. LIKERT TESTING

Likert testing is widely used by researchers. It is a test of a person's subjective response to a statement. Most often the test asks how strongly a person agrees or disagrees with a particular statement put to the person. A common form allows participants to select between five possible responses: 'strongly agree', 'agree', 'neither', 'disagree' and 'strongly disagree'.(Heiberger and Robbins, 2013) Figure 2 is an example of a likert test presented to participants during our study.

Likert testing has been applied to readability studies in previous research. Heydari employs likert testing to evaluate the readability of ten passages.(Heydari and Riazi, 2012) Hall and Hanna use likert testing to assess the effect of colour on readability of web pages.(Hall and Hanna, 2004) Ferrari and Short apply likert testing to evaluate the effect of size and font type on readability.(Ferrari, 2002)

Kandula et al use seven point scale likert questions with a cohort of experts and patients to rate the readability of health literature. They are concerned with the difficulty of health literature which they note

the Institute of Medicine assessed was difficult to read or act on by more than half of the US adult population. They found a high level of correlation (.81) between expert and patient ratings of language difficulty.(Kandula and Zeng-Treitler, 2008)

The appropriate analysis of likert scale items is a matter of controversy among researchers and the question is relevant to analysis of our results. Different camps argue for different analysis methods. Essentially, the controversy concerns whether parametric as opposed to non-parametric tests can be used to analyse likert data.[11] Clason et al argue that likert items must always be treated as ordinal, even when combined in a scale, and therefore argue for non-parametric testing. (Clason and Dormody, 1994) Norman critiques such arguments, arguing that parametric tests are often robust even when assumptions (such as normality) are violated. *"[B]oth theory and data converge on the conclusion that parametric methods examining differences between means, for sample sizes greater than 5, do not require the assumption of normality, and will yield nearly correct answers even for manifestly non-normal and asymmetric distributions like exponentials."* Norman concludes that parametric tests are appropriate for analysis of likert data both for differences of means and correlation of data.(Norman, 2010) Similarly de Winters et al, who undertook a systematic comparison of t-tests (a parametric test) and the Mann Witney Wilcoxon test (a non-parametric test) on a diverse range of distributions of data concluded that the differences between the tests was minor and exceeded 10% only for a few of the 98 distributions they studied.(de Winter and Dodou, 2010) To the extent that parametric analysis is used on likert tests in this paper, there is sufficient support for it in the research literature.

## 2.8. Cloze testing

The cloze procedure involves testing the ability of readers to correctly re-insert words that have been deleted from a given text. Typically the test is administered by deleting every $n$th word in a text. When used to assess the readability of a text, the cloze procedure is administered by deleting every fifth word (including sometimes five different versions of the text staggering the deletion), and replacing it with a blank space. The reader must fill in the missing terms.(Bormuth, 1967) Figure 3 is an example of a cloze test used in our research.

Although initially conceived as a remedy for the shortcomings of readability formulas, the cloze procedure came to complement conven-

---

[11] Parametric tests (such as the students t-test and ANOVA testing) make more assumptions about the test data than do non-parametric tests.

Citizen Science for Citizen Access

tional reading tests.(DuBay, 2004) Cloze procedure was also developed to provide a more valid measure of comprehension than traditional multiple choice comprehension tests.(Wagner, 1986) Of greatest interest in this context is use of cloze tests as a measure of the readability of a text. Bormuth notes that there is a high correlation between cloze readability testing and comprehension testing on human subjects:

> *A reasonably substantial amount of research has accumulated showing that cloze readability test difficulties correspond closely to the difficulties of passages measured by other methods.* (Bormuth, 1967)

Bormuth cites studies, including his own, which show correlations ranged from .91 to .96 with the difficulty of texts assessed with traditional comprehension tests.(Bormuth, 1967) When properly applied, the cloze test provides an indicator of how difficult a text was for given readers. A cloze score (i.e. proportion of correct responses) below 35% indicates reader frustration, between 35% and 49% is 'instructional' (the reader requires assistance to comprehend the material) and 50% or above indicates independent reader comprehension.(Wagner, 1986)

## 2.9. SEMANTIC DIFFERENTIALS

Semantic differentials were originally developed by Osgood in the 1950's. A semantic differential is comprised of two bipolar adjectives ('readable-unreadable' for example) with a scale in between. The research participant is asked to select a point on the scale which they consider best corresponds to the test stimulus. Typically, the user is presented with multiple semantic pairs and asked to assess a test item for each pair. Figure 4 provides an example of a semantic differential test used in our study.

Semantic differentials may vary by number of points on the scale or presence or form of labelling of scale point. A scale varies from a positive to negative end and thus has both direction and magnitude. (Garland, 1990; Johnson, 2012) Semantic differentials have been widely applied and are seen as an accurate measure of individuals 'affective' responses to a stimulus. Osgood found that users ratings of semantic differentials could be reliably grouped into three major dimensions which he labelled evaluation, potency and action. The method has been used to test individuals responses to words, pictures, facial expressions and a wide variety of concepts.(Johnson, 2012)

Garland compares three different forms of the semantic differential test to test whether the form of the test affects user responses. The three forms were: semantic differentials without labels, semantic differentials with numeric labels and semantic differentials with text labels (such as 'very', 'quite', 'neither' etc). Garland asked users to rate the test for

preference, ease of expressing opinion and ease of completion, finding that users preferred semantic differentials with text labelling. Garland also found that there was no difference in the distribution of responses of the administered semantic differential tests and concludes that the form of test used is unlikely to influence users responses. Garland does however note that numerical scales may be favoured by users who are used to working with numbers. (Garland, 1990) For these reasons, and given that 'used to using numbers' is not a characteristic applying particularly to the users of legislation we use a labelled semantic differential test in our study.

Semantic differentials have also been used to measure "user experience". User experience has been defined as "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service". The concept is broader than concepts of usability which are more specifically concerned with functional characteristics of the artefact being tested. (Vermeeren et al., 2010) The use of semantic differentials in our study enabled a broader examination of how users responded - for example including responses to concepts such as 'leniency' or 'attraction-repulsion'.

A possible alternative to a semantic differential (in our case participants were asked to select an appropriate radio button) is the 'visual analog scale' or a slider. However, Couper et al find no advantages to use of a visual analog scale. Rather they found that using a slider led to higher levels of missing data, and longer completion times.(Couper et al., 2006)

## 2.10. Principal components analysis & factor analysis

In our study we collect not only multiple ratings, but also multiple ratings of multiple variables for each sentence used in the study. We need to combine the data from each of cloze, likert and semantic differential test to provide a single variable which is representative of reading difficulty of a particular sentence.

Principal Components Analysis is particularly suited to this task. Its goal is to reduce the number of dimensions in a set of observations by combining variables into a reduced number of variables.(Härdle and Simar, 2003, p 234, 241) Factor analysis similarly seeks to identify underlying latent variables (factors) by grouping together variance in the most highly correlated variables into a reduced number of factors.(Floyd and Widaman, 1995) A question for both methods is how to decide how many variables to retain after principal components or factors are extracted. One widely supported method is graphical. It looks for a bend in a curve known as a scree plot. The y-axis of the

Citizen Science for Citizen Access

scree plot shows eigenvalues extracted for each principal component, while the x-axis shows the extracted components themselves. Figure 16 is an example of a scree plot. Principal components to the left of the 'bend' in the scree plot are often retained. This method is regarded as among the most sound.(Costello and Osborne, 2005)

## 3. Description of the Study and Observations
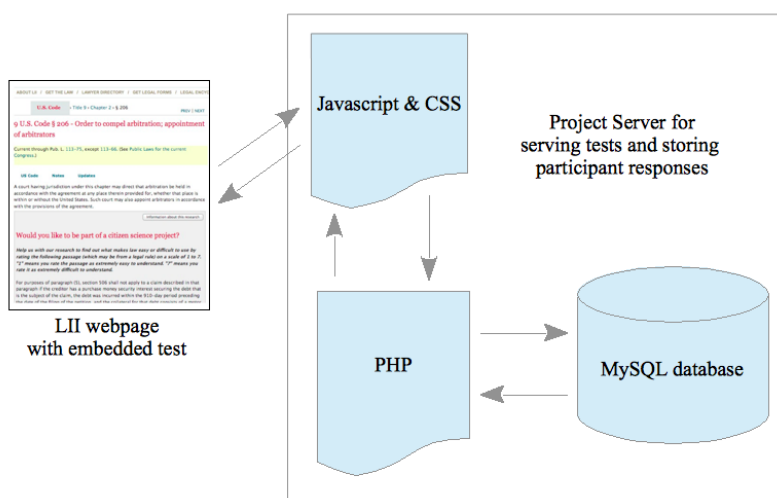


*Figure 1.* Project platform design

The primary data for the research was collected from 2 May 2014 until 31 July 2014 using the crowdsourcing methods described above. In total, 63,250 submissions were received from users spread across four sets of data: demographic data, likert submissions, cloze results and semantic differentials. From among these, some submissions were null results (e.g. a user pressed the submit button without providing any data). Also some data was removed as outliers for particular tests.[12] Table I below shows totals and percentages of usable data after filtering. More than 43,355 usable readability assessments were collected.

---

[12] For semantic differential tests, 'donkey votes' were removed - i.e. votes in which the user selected only the same value down a column. Also results where more than 30% of a semantic differential were null were removed. For cloze testing, an issue where a score of 0 was obtained, was how to distinguish genuine attempts from 'careless' input. Results were filtered if 30% or less of fields of a cloze submission had any input (i.e. an attempt at guessing the word). As noted above, the issue of data validation is a characteristic of citizen science projects.

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

Table I. Submissions

| Database | Total Submits | % of Usable Data |
|---|---|---|
| Demographic Data | 14912 | > 93.8 |
| Likert Data | 23402 | 99.9 |
| Cloze Data | 12970 | 85.2 |
| Semantic Differential Data | 11966 | > 74.6 |

Over the period of three months that data was collected, the rate of data collection remained essentially linear. This points to the feasibility of longer term data collection of user experience in online legal publishing environments. Also, as there was an equal chance of being asked to complete a likert, cloze or semantic differential test, the response rate for each question type is informative. Semantic differential tests were least likely to be responded to, while likert tests were responded to at almost twice the rate. As rate of data collection was an important consideration, these differences are relevant to future research design.

The tools used to undertake the research included the python programming language, used for scripts for preparing corpora and undertaking data cleaning extraction, the Weka Data Mining Software package (Hall et al., 2009) for machine learning, the R-statistical package and associated R-Cmdr graphical user interface (R-Core-Team et al., 2012; Fox, 2005) for undertaking statistical analysis, the Readability Research Platform (Curtotti and McCreath, 2013) and the Natural Language Toolkit (Bird et al., 2009) for carrying out metrics extraction and natural language processing.

The platform to enable data collection had a number of functional elements:

**(a)** a background server for serving test sentences and receiving and storing participant responses;

**(b)** php scripts which communicated asynchronously with a mysql database on the server;

**(c)** javascript and css files which communicated with the primary Cornell pages and with the php scripts.

Figure 1 provides an overview of the platform design. Brief code snippets in the primary LII pages linked the platform with the web pages viewed by participants.

Figure 2 illustrates likert tests used in our study. The participant was presented with a sentence selected from the four test corpora and

Citizen Science for Citizen Access



*Figure 2.* An example likert test presented to research participants

was asked to indicate their level of agreement from "strongly agree" to "strongly disagree" with a statement about a sentence. That statement could be that the sentence was "very easy", "easy", "hard" or "very hard" to read. For example the user might be presented with the statement that "The text is very easy to read" and asked to indicate their level of agreement with the statement.

Figure 3 illustrates a cloze test. In this case, the participant was asked to guess up to ten missing terms in a test sentence.

Figure 4 illustrates a semantic differential test. Here, the participant was asked to rate a test sentence against each of ten semantic differentials. As we wished to minimise potential disruption to the normal use of the LII website, tests were presented at the bottom of LII pages.

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce



*Figure 3.* An example cloze test presented to research participants

## 4. Demographics

The demographic results from our research is a particular focus of this paper. We first present Google Analytics from the LII site which provides an independent source of data addressing user behaviour on the site. We then discuss the demographic data collected during our study.

### 4.1. WHAT LAW DO PEOPLE READ? INSIGHTS FROM GOOGLE ANALYTICS

Google Analytics were studied for visits on the Cornell LII legislation pages over a period of 12 months. In the period 18 October 2012 to 17 October 2013 a total of 927.4 person years were spent reading legal rules at the LII site (this includes the US Code, CFR, UCC, constitution, rules of procedure etc). Most people found their way to legislative provisions directly by searching for the relevant legal rule (i.e. the landing page on the LII server was a specific section or regulation). This implies that often people have had some introduction to what might be relevant

Citizen Science for Citizen Access



*Figure 4.* An example semantic differential test presented to research participants

laws for their concerns, before navigating to the LII site. The site has a large traffic, with 112 million page views during a year (of which 38.8 million page views are from US Code and 19.5 million page views are from the Code of Federal Regulations) (21 August 2013 - 20 August 2014). By comparison the official UK legislative site receives 5 million page views per week.(Tullo, 2013)

Most interesting from the readership data was its power law distribution. Far from readership of sections being equally distributed - the readers for a particular section might attract varied by many orders of magnitude. A mere 37 sections of the US Code (landing pages), for example, account for 9.97% of entire traffic to US Code sections.
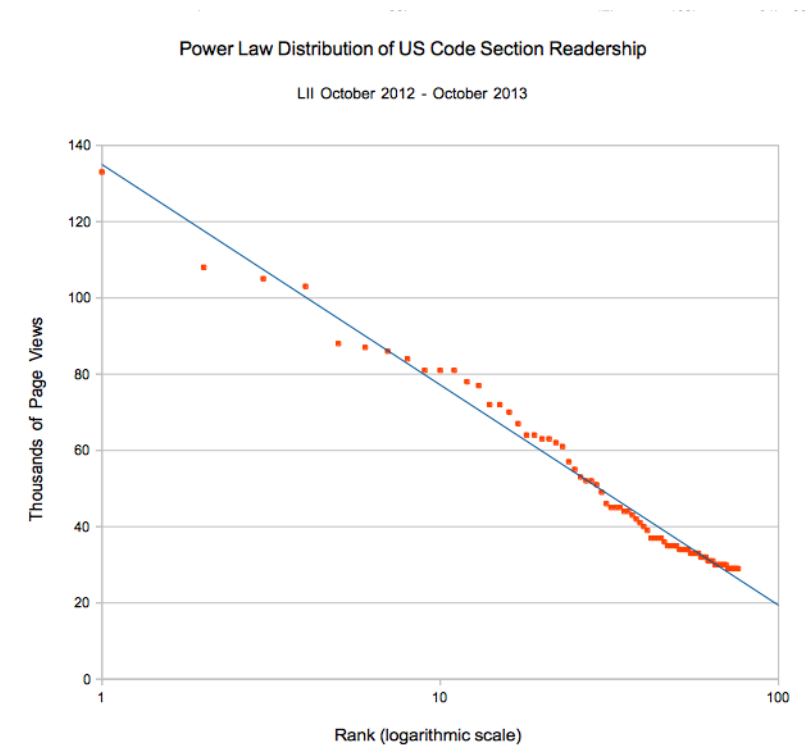
Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

Power Law Distribution of US Code Section Readership

LII October 2012 - October 2013



*Figure 5.* Power law distribution of section readership for most frequently visited sections

This was from a total 'node' count of close to 65,000 sections. Of these, 8391 sections (pages) were visited once in a 12 month period, 4267 were visited twice, and 2833 were visited 3 times. The most frequently visited section was visited 133 438 times during the twelve month period (Title 28 section 1332).

The implications are significant for the task of enhancing access to law. For practical purposes, most of the US code is of marginal relevance. It is rarely read and efforts to improve its readability may not be warranted. On the other hand, language difficulty in highly read parts of the code will impact significantly on access to law and on the regulatory burden faced by users. For a much smaller effort than full review of an area of law, a disproportionate improvement in user experience is available by addressing readability of the most read parts of the code.

Citizen Science for Citizen Access

## 4.2. Demographic data

As mentioned above, research participants were asked to provide general demographic information, but could opt out if they wished to do so. The population from which demographic data (and other data) was drawn was limited to visitors to the site who were engaging with legislative or regulatory materials (i.e. legal rules). Data was collected on age, birthplace, education, gender, language and persona. The 'persona' category refers to certain typical users of legal data: e.g. legal professionals and members of the public.

There may be systematic effects in those who voluntarily chose to participate in the study, nonetheless the results provide an indicator of the user base for online legislative information.

On questions of readability, Dubay notes the two most important questions are "the reading skills of the audience" and the "readability of the text".(DuBay, 2004) For example, if all readers of law are judges (i.e. an audience highly familiar with reading and comprehending legislative texts) readability issues will play out quite differently to a case where a substantial proportion of readers are not legally trained. In the latter case, such readers may find the language difficult and unfamiliar, and a case may be established for improving the writing of legislative texts.

## 4.3. Who reads the law online and why they do so

Participants were asked to nominate a broad persona that best matched the reason they used the law. The use of personas to study readability in a legislative context was described in a study of the users of UK Legislation reported by Carol Tullo of the UK National Archives Office at 2013 Law Via the Internet Conference. In the case of that research, the personas were: a compliance officer; a law librarian; a member of the public seeking to defend her rights; and a parliamentarian. It was noted that such categories do not necessarily capture the entire user base.(Tullo, 2013)

In our study five personas were used: legal professionals (including law students); non-lawyers engaged in compliance; members of the public seeking information on their rights; individuals engaged in law reform or law making; and "others". As would be anticipated legal professionals and legal students (i.e. the legally trained) were the largest single group of respondents (41.7%). However, surprisingly, they were the minority of respondents.

Members of the public seeking information on their rights (23%) and non-lawyers engaged in compliance management (13.4%) also represented substantial categories. A large "other" category represented

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

(18.59%) responses. This category was almost one fifth of respondents, although it is not immediately obvious what this 'other' category may represent. Those engaged in reforming the law (participants relevant to the democratic process) represent 3.5%. (See Figure 6) Meeting the needs of users drawn from the public is most directly related to "access" for reasons of equity. The compliance category represents considerations of economic efficiency. The reform category is related to rule of law and the democratic process.

## Persona
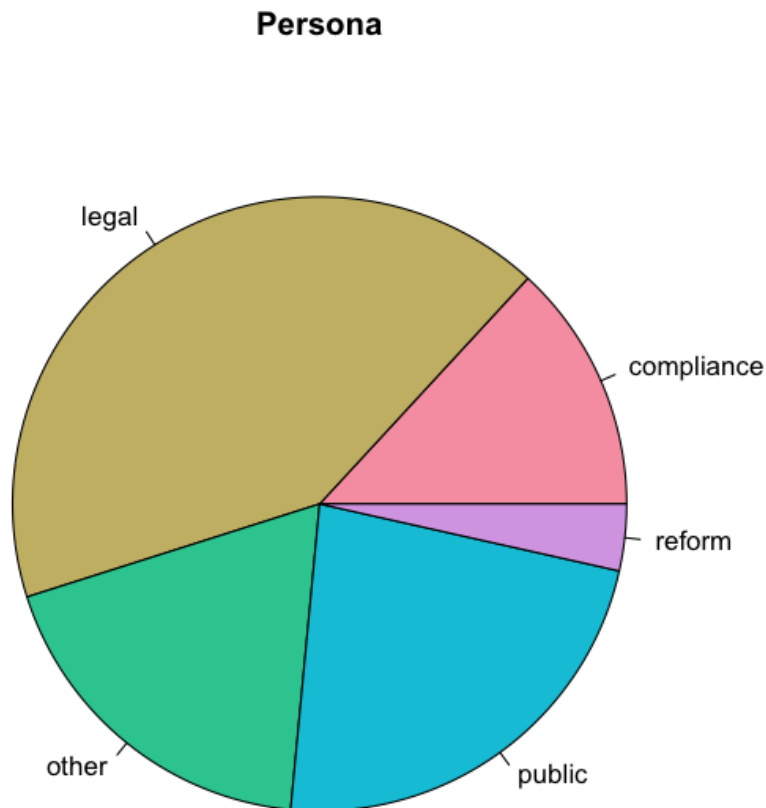


*Figure 6.* Who reads legislation

We also explored differences in *which parts* of the law different user categories were most likely to be reading. This data was derived from the landing or source page from which a reader participated in the research.

Lawyers were far more strongly represented in the audience for Federal Rules as opposed to the other two bodies of legislation (66% versus around 40% for the US Code or Code of Federal Regulations).

Citizen Science for Citizen Access

Those concerned with compliance represented 21% of the audience for the Code of Federal Regulations, dropping to 9% of the audience for the US Code. By contrast the public audience for the US Code was 27% with the 'other' group representing another 20%. Table II sets out additional results.

Table II. Percentage of Audience By Legal Code

| code | compliance | legal | other | public | reform | count |
|------|-----------|-------|-------|--------|--------|-------|
| Code of Federal Regulations | 21.3 | 40.4 | 17.7 | 17.5 | 3.2 | 5119 |
| Federal Rules | 3.3 | 66.7 | 8.2 | 18.0 | 3.9 | 672 |
| US Code | 8.9 | 40.6 | 20.0 | 26.8 | 3.7 | 8188 |

The top eleven most frequent titles of the US Code were also examined for distribution of audience. Figure 7 illustrates the distribution of audiences. Legal professionals constituted a majority of the audience only in the cases of Title 28 (the judiciary), Title 11 (bankruptcy) and Title 36 (patriotic observances). The public were the highest users of Title 18 (crime) and Title 10 (armed forces). Individuals interested in legal reform were most highly represented in Title 17 (copyright), although in no case representing more than a few percent of total audience. Compliance officers were most highly represented in Title 26 (internal revenue). In terms of access again implications can be drawn. Criminal law is an area where the public needs the law to be readable. They are a substantial audience for the criminal law. It will be noted from Figure 7 that it is a heavily read title. The internal revenue code (unsurprisingly) represents an area where concerns relating to regulatory burden are more pertinent.

The results discussed above are consistent with the observations of the UK Parliamentary Counsel's office that *"increasingly, legislation is being searched for, read and used by a broad range of people…; websites like legislation.gov.uk have made it accessible to everyone."* (OPC-UK, 2013) The consistency of our results with the description of the user base for a major national legislative site support a conclusion that the patterns observed on the LII site are not an artefact of either the LII site or the study design. The results support a conclusion that whatever may have been the situation in the past, legal professionals are far from the primary readers of legislation in the online environment. Indeed substantial non-lawyers are a substantial audience for the law online. Some caution is required in interpreting this result, as it is possible that other reasons explain it (e.g. lawyers might have responded at a lower rate than non-lawyers). Studies on other online sites would be required

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce
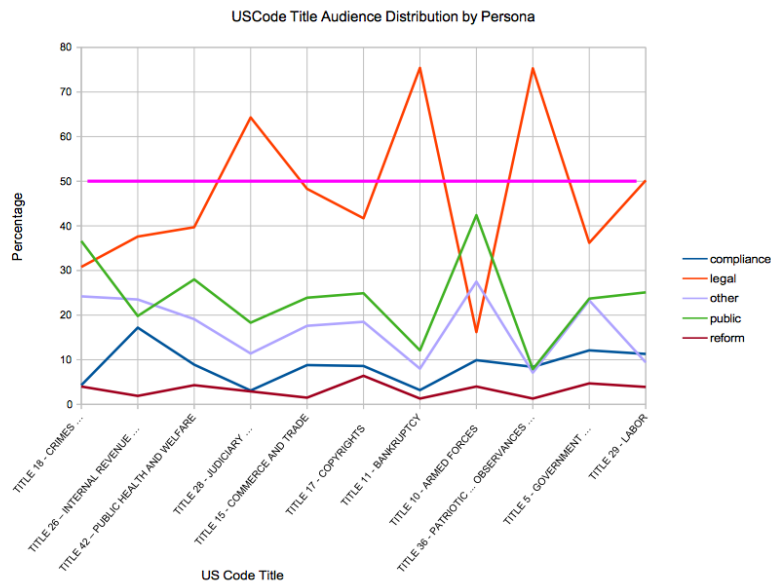


*Figure 7.* Persona distribution by title

to clarify whether this result generalizes to the underlying audience for law online.



*Figure 8.* Relative readership of US Code titles

In light of such findings as to the user base for legal rules, "who" legal rules are being written for is practically as well as theoretically important. To write only, or primarily, for judges and lawyers fails to address the needs of a substantial proportion of users of legal rules. Also it is possible to differentiate between different parts of legislation

Citizen Science for Citizen Access

by level of audience interest, again carrying implications for how law might be written.

## 4.4. GENDER RESULTS

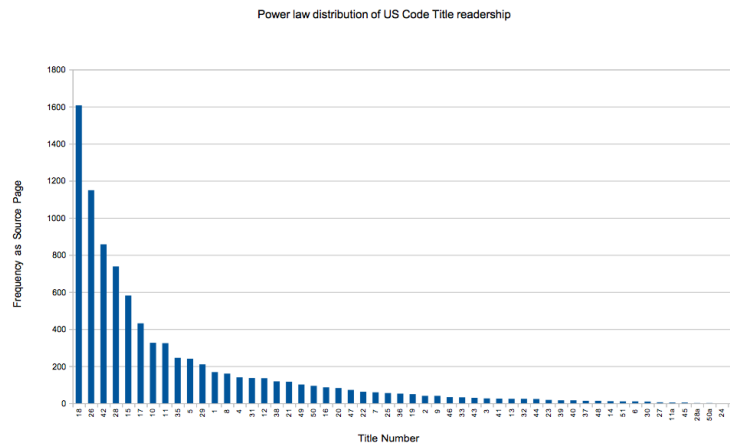There was a sharp disparity in the number of responses received from males and females. Women represented only 35.4% of the responses by gender. There are a number of possible explanations for this result. One possibility is that there is a gender disparity in access to law reflecting societal conditions. There may be factors in the way that law is provided online that affects its accessibility to women. Alternatively, the result may be wholly or partially an artefact of the study design. To promote participation, the study was described as an invitation to participate in 'citizen science'. If the 'scientific' description is a cause of the lower participation by women, it is a marker of gender exclusion in another social dimension. A further possibility is that the results at this site are not representative of broader usage patterns.

The legal profession, like many professions has only partially achieved gender equality. It would be expected therefore that there would be a lower representation of women in the legal profession persona and this is reflected in our results (60% to 40%). However, the gender disparity in participation is even more marked for the non-professional personas (i.e. public and other), where women represented only 30.3% and 33.5%, respectively. The differences between these different personas tend to support a conclusion that the difference in participation is 'real', rather than related to the study design. Such gender disparities merit further investigation.

## 4.5. AGE

Users were asked to nominate an age category (grouped into 15 year age bands). The responses show a broad distribution across age groups. Figure 9 illustrates these results. However age is not evenly distributed across user groups. Legal professionals are dominated (as would be expected) by working life adults. They are also generally younger as compared with members of the public accessing law online. This may reflect the inclusion of law students in this group.

## 4.6. BIRTHPLACE

Users were asked to nominate a broad region of the world in which they were born. Over 85% were born in the United States or Canada. This is to be expected for a US legislative site. The overseas born population reported in 2010 for US population was 12.9%.(Grieco et al.,
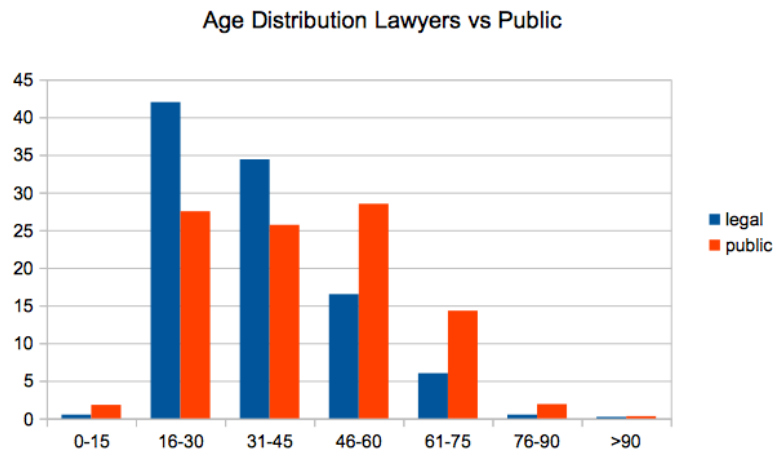
Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

## Age Distribution Lawyers vs Public



*Figure 9.* Legal vs public age distribution

2012) This is a similar proportion to the proportion accessing the LII legislative pages, although direct parallels are invalid, because LII users also include an unknown number of users from overseas.

### 4.7. EDUCATION

Users were asked to indicate the highest level of education they had completed: primary, secondary, vocational or tertiary. Overall tertiary respondents represented 78.24% of respondents. Primary were 3.74%, secondary were 9.31% and the vocationally educated were 8.72%.

As with age, educational completion varied significantly between different personas. The tertiary educated strongly dominated the legal profession (94.3%), as would be expected. For the public, the proportion of tertiary educated was 56.9%. This figure is considerably higher than the completion of tertiary education in the US population as a whole. Again this suggests an access issue. Those with primary, secondary and vocational education are under-represented among readers of law. In 2009, between 20% and 30% of the population over age 25 had completed tertiary education.(Ryan and Siebens, 2012)

Again this carries implication for access to law. Many without tertiary education may not even be attempting to read the law. In addition, 43.1% of those who were among participants did not have tertiary education. To address the needs of the public, the law needs to be written to take account of the fact that a substantial proportion of its readership does not have tertiary education. In terms of the regulatory burden, of those concerned with compliance, 16.6% do not have tertiary
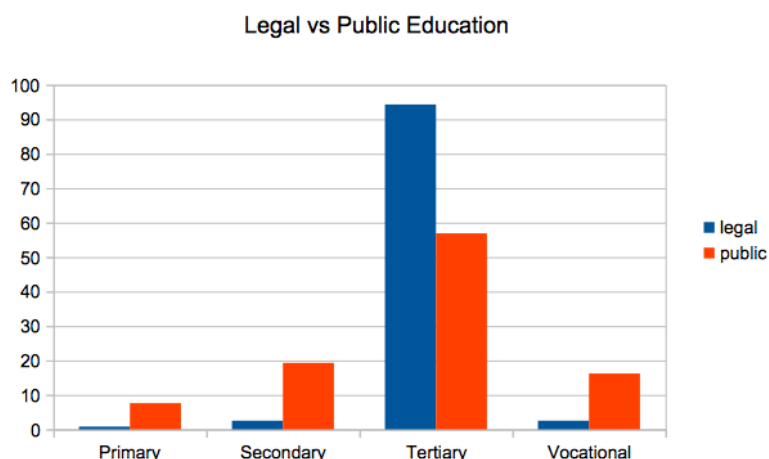
Citizen Science for Citizen Access

**Legal vs Public Education**



*Figure 10.* Legal vs public educational attainment

education. Further consideration is required of the actual educational attainment of the population as a whole, which suggests that for law to be more accessible to a larger proportion of those who read (or might read it in future), the law should be designed to be readable to those whose education is limited to secondary.

### 4.8. Language

Users of the site were asked to identify the language they spoke best. 93.6% identified English as their primary language. 2.01% identified Spanish, while 4.36% nominated 'other language' as their primary language. In the US, the population is primarily English speaking. However in 2011, 37.6 million people in the US spoke Spanish at home (about 12.9% of the population as a whole). Of these 25.9% self-identified as not speaking English well. The usage of the site by language may suggest lower access to law for the US population which is primarily Spanish speaking. This issue also has geographical implications as the Spanish speaking population is not uniformly distributed throughout the US, but is particularly concentrated in western and southern US states.(Ryan, 2013)

### 4.9. How does reading difficulty vary by demographic groups?

Although three readability datasets were collected, cloze results are most reliable as a measure of reading difficulty when used to draw comparisons between different demographic groups. Cloze tests, in contrast

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

to likert and semantic differential tests, provide an objective measure of reading difficulty for a reader. Likert tests and other subjective responses are affected by user perceptions and background as well as the test stimulus which may produce unreliable results when comparing between different demographic groups.

The results below are based on raw correct score results for cloze tests broken down by demographic groups (rather than proportion of correct scores).

### 4.9.1. *Cloze results by persona*

Figure 11 shows average correct scores for different personas for different corpora. The 95% confidence interval of the mean is also shown.[13]
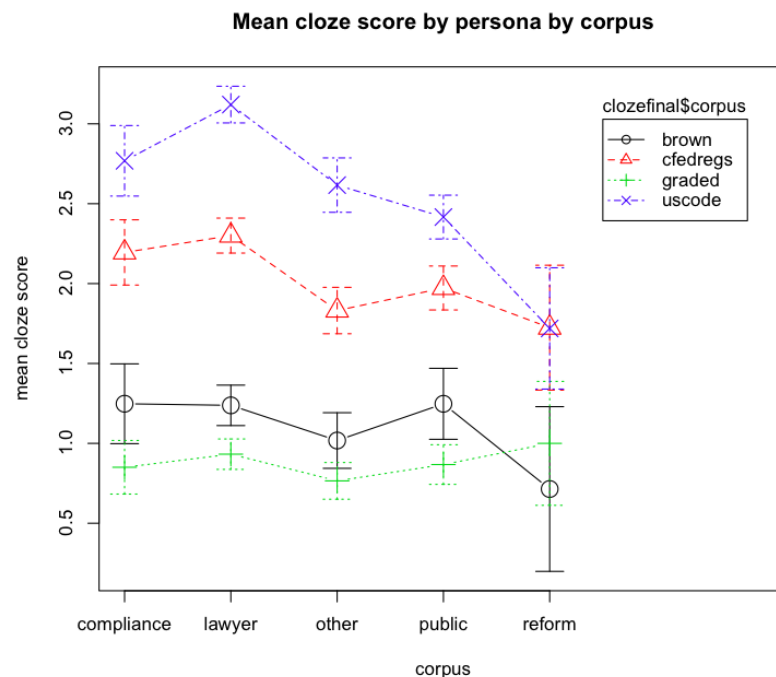


*Figure 11.* Mean cloze correct with 95% confidence intervals by persona by corpus

By visual inspection we can see that mean results for lawyers and the public and 'other' groups are significantly different for sentences from the US code and the Code of Federal Regulations. It is also notable that the demographic differences for the graded sentences and sentences

---

[13] Note that Figure 11 cannot be read to compare the corpora against each other. This is because each corpus has a different distribution of sentence lengths and the cloze scores are dependent on sentence length (e.g. there is a higher proportion of short sentences in the graded corpus than in the US Code).

Citizen Science for Citizen Access

from the Brown corpus are not significant. Figure 12 shows confidence intervals by persona for the US Code. This last diagram allows us to conclude that lawyers do better than all groups, including the reform or democratic element. The reform group showed the greatest difference in means with the legal group.

**95% family-wise confidence level**
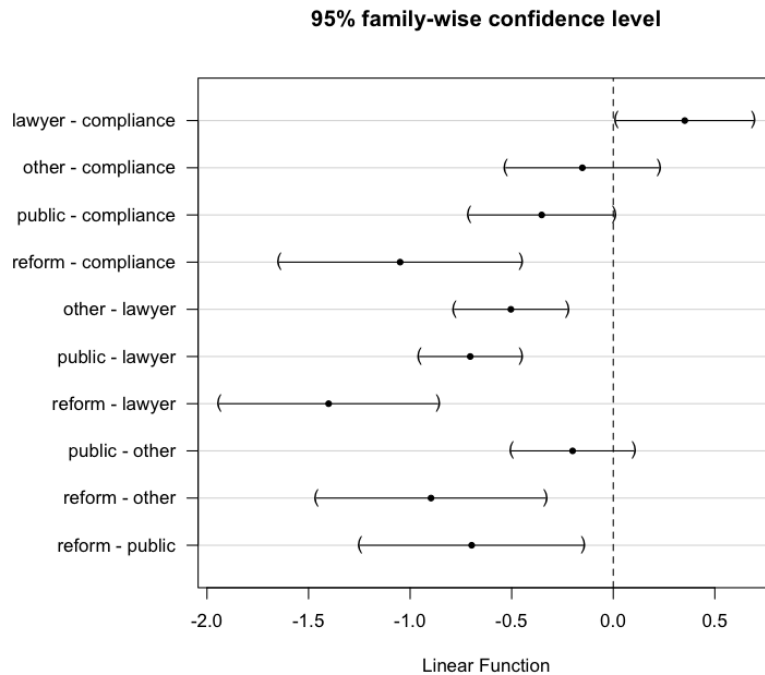


*Figure 12.* Mean cloze 95% confidence intervals by persona for the US Code

A one-way ANOVA test was also carried out to test for significance of differences on the mean score for the US Code for different personas finding that the mean differed significantly by persona, $F_{(4,3049)} = 23.47$, $p =< 2e-16$ with effect size 0.023 (i.e. small).

To use the cloze tests to measure language difficulty as per the standard cloze readability methods developed by Bormuth and others (see Section 2.8 above) we need to calculate the proportion of correct responses for cloze tests of the same length. This can be done by taking a subset of data - e.g. the data in which all tests have ten gaps. Such a filtered dataset was prepared. Also sentences from the Brown corpus were removed from this subset to provide a set of cloze tests solely on sentences from the two legal corpora.

This dataset consisted of 2556 cloze test responses. Lawyers achieved average cloze proportal score of 0.42 while members of the public, compliance, and democratic groups achieved 0.35, 0.39 and 0.26, respec-

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

tively. There is a significant difference of means between lawyers and other groups (at $p < 0.001$) in all cases except for the difference between lawyers and the compliance group (which was not significant). Interestingly, those involved in the democratic process achieve the lowest proportional results. In Section 2.8, it was noted that results between 0.35 and 0.49 indicate the reader needs assistance to comprehend the material. Results lower than 0.35 indicate the reader is frustrated by the material. Our results, which point to legal materials being very hard to incomprehensible for many audiences, are consistent with studies described in Section 2.3 which discuss the readability or otherwise of legislation.

### 4.9.2. *Cloze results by other demographics*
For reasons of space, differences in readability difficulty for other demographic groups are not discussed. However the following is interesting to note. On average, women obtained a higher average cloze result than men. This was largely a result of women performing significantly better than men on cloze tests on the graded corpus.

### 4.10. Subjectivity, likert results and semantic differentials

Although useful for between sentence comparisons, likert results are subject to a number of issues when used for comparison across different demographic groups. For example the desire to agree with the questioner varies between cultural and other groups. In our study this effect can be very clearly seen as between Spanish speaking respondents and other language groups. Figure 13 shows level of 'agreement' by likert question. The x-axis shows the question type (i.e. whether the test sentence was easy, hard, very easy or very hard to read). The y-axis shows mean response for each question type by corpus. A lower mean indicates a higher level of average agreement. The graph shows that Spanish speakers are more likely to agree with the questioner, irrespective of question asked. Accordingly comparisons between demographic groups need to be approached with caution.

Semantic differentials are also potentially affected by the subjectivity of individual responses. For example average results for semantic differentials for gender show that women rate sentences as less readable than men, yet we saw above that in terms of mean cloze results, women scored higher than men.

Note that neither likert nor semantic differentials showed demographic differentiations in readability difficulty between demographic groups that are evident with the cloze test results. This 'subjective'

Citizen Science for Citizen Access
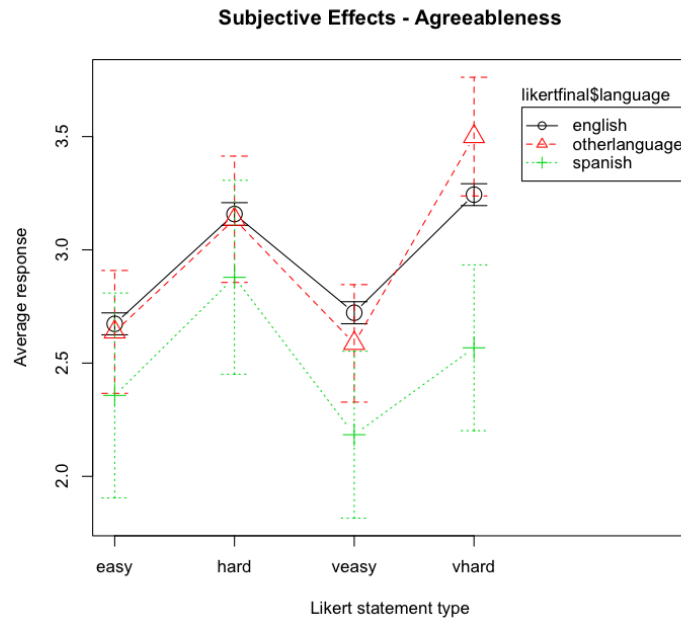


*Figure 13.* Subjectivity effects between demographic groups

result is interesting to compare with the study by Kandula et al. and de Clercq et al., which compared expert and non-expert evaluation of sentences. Both studies found that experts and non-experts tended to have a high level of agreement when asked to give their subjective judgement as to difficulty.(Kandula and Zeng-Treitler, 2008; De Clercq et al., 2013)

## 5. Measuring the Difficulty of Sentences

In this section, we turn to the question of measuring the difficulty of sentences. The data collection phase provided three datasets that could potentially contribute to the development of a ranking of sentences by difficulty (the likert, cloze and semantic differential datasets). We systematically examine each of the datasets. We also consider how the results from each set can be combined into a final measure.

### 5.1. Likert results

The likert dataset consists in reality of four sub-sets of data depending on the question that the user was asked.

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

The likert dataset was the largest, as it was most often responded to by research participants. On average, 17.86 responses were provided for each sentence, with a standard deviation of 4.39. The distribution of the number of responses by sentence was approximately normal.

Figure 14 shows the distribution of degree of "agreement" depending on the question the participant was asked. The x-axis represents degree of agreement, with 1 being 'strongly agree' and 5 being 'strongly disagree'. The y-axis peaks show proportion of responses by question type.[14] Broadly, the distribution of responses between easy questions and hard questions mirror each other. This is consistent with our intuitions as to how meaningful responses should be distributed.
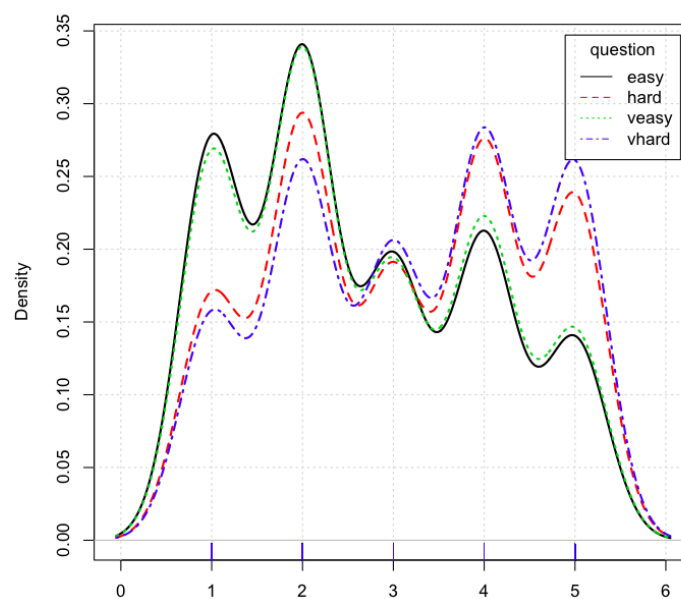


*Figure 14.* Density distribution of responses by question

It is also helpful to visualize average level of agreement by question and corpus, together with their 95% confidence intervals. Figure 15 provides basic validation that user responses can be used to distinguish sentences by level of reading difficulty. i.e. the averages for corpora are consistent with our expectations for the difficulty of each corpus. It can also be seen from the confidence intervals of the means that in all

---

[14] This graph was produced using the R statistical package. For visualizing the bandwidth (width of waves) has been artificially increased to aid visualisation. Note that 'neutral' responses and 'not sure/not applicable' responses have been combined in our analysis.

Citizen Science for Citizen Access

cases there is a significant difference of means between the legal and non-legal corpora, except for the "very easy" question, which in the case of the difference between the Brown Corpus and code of federal regulations did not show a significant difference at the 95% confidence level.
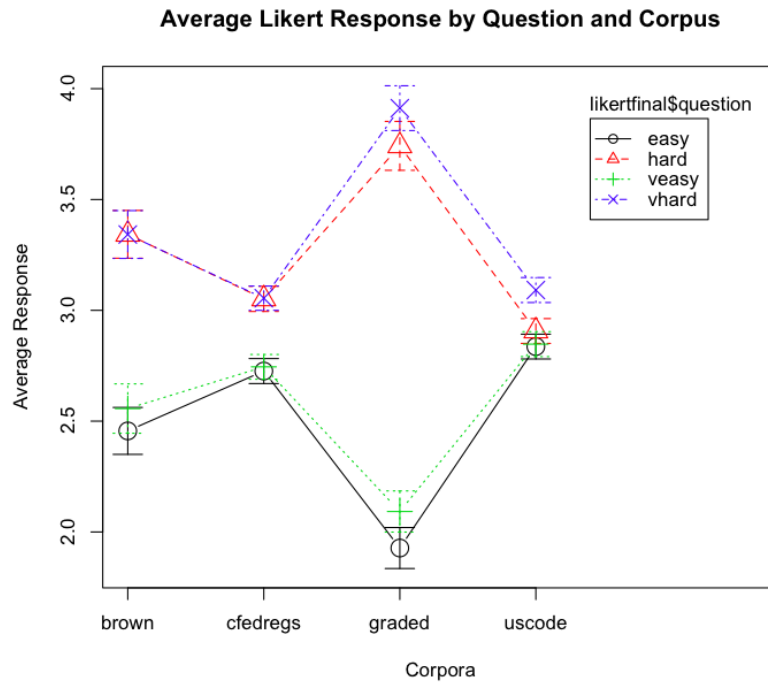


*Figure 15.* "Mean" likert response by question and corpus with 95% confidence intervals of the mean

To assign a composite likert measure of reading difficulty to each sentence by combining the results of the various responses, principal components analysis was applied. (See discussion above in Section 2.10) The input variables were the proportion of responses for each category in the likert test (i.e. the 20 categories (5 possible responses x 4 question types). After extracting the principal components we examined a scree plot for the data. This identified the first principal component (i.e. the component before the bend in the scree plot) as sufficient to represent the variance in the data. (See figure 16)

As a sanity check, this first principal component was compared with an aggregate measure derived by calculating the proportion of 'votes' from users indicating a sentence was 'hard' less the proportion of 'votes' that a sentence was 'easy'. This was done by first recoding

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

**.PC**
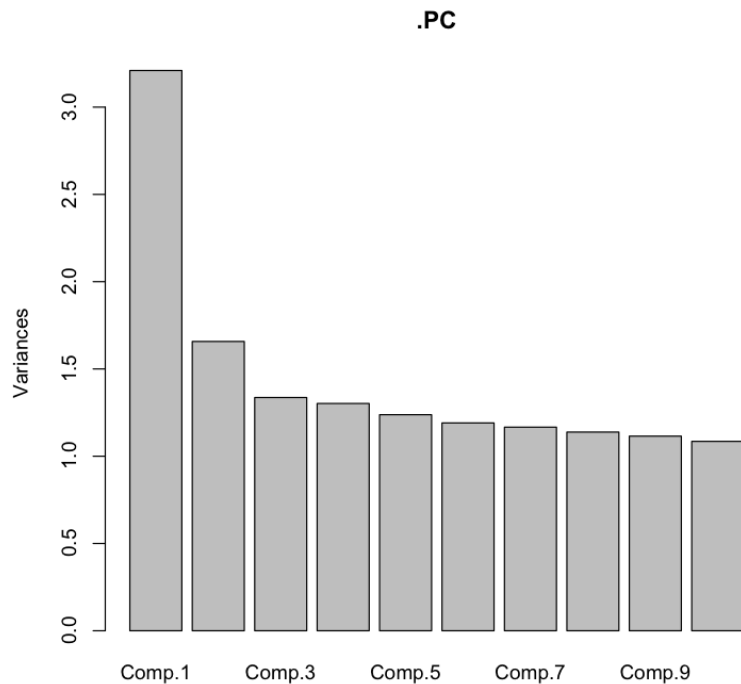


*Figure 16.* Scree plot of principal components extracted from likert data

and binning each response into a "hard", "easy" or "neither" vote and then calculating the proportion of votes cast for a sentence in each bin calculated. For example 'strong agreement' that a sentence is hard is classified as a 'hard' vote; and 'strong disagreement' that a vote is easy is also classified as a 'hard' vote. This 'hard-easy' variable is highly correlated with the proportion of hard votes (at 0.96) and the proportion of easy votes (at -0.97). It is also correlated at 0.95 level with the first principal component described above. Notably, the second and third principal components had low correlation with any of these measures (the highest correlation being 0.22). Principal component 3 did however correlate at -0.72 with the proportion of 'neither' votes (i.e. votes where a participant did not indicate that the sentence was either hard or easy).

We further explored the first principal component by breaking down the data into the four corpora. Figure 17 illustrates the distribution of the first principal component for sentences for each of the four corpora.

Citizen Science for Citizen Access

It will be evident that the metric is broadly normal and that the four corpora have different means.
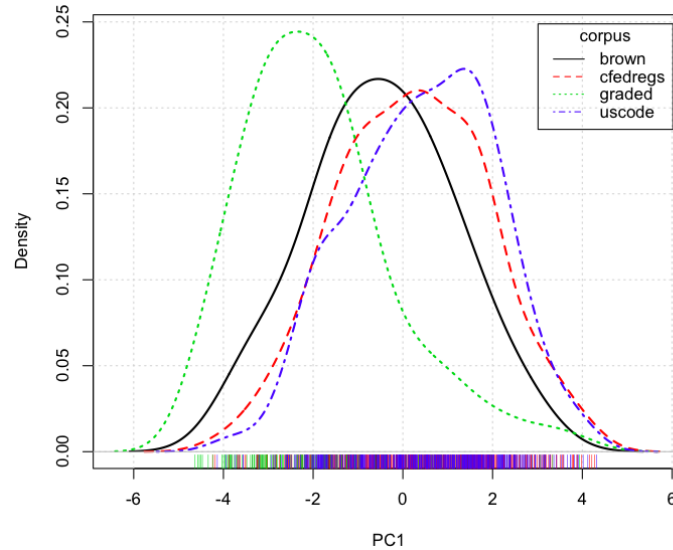


*Figure 17.* Density distribution of first principal component from likert results by corpus

ANOVA testing on differences between the mean results of the first principal component was carried out, as well as pairwise mean comparison. ANOVA returned a significant difference as did a comparison between all corpora (at $p < 0.001$) except between the US Code and the Code of Federal Regulations.

It is also worth noting the overlapping distributions of the corpora. Language is not sharply delineated into 'hard' and 'easy' categories: rather each sentence falls on a continuum of difficulty. This is relevant to the task of sentence classification used in machine learning which by its nature requires data to be assigned to categories. In reality reading difficulty does not come in neat separate packages that are easily detected. Most sentences are found close to a mean readability value.

5.2. CLOZE RESULTS

We now turn to an analysis of the cloze results, similarly for ranking sentences by language difficulty and assigning a difficulty level to each sentence.

Sentence length strongly affects the cloze results for each corpus. This is due to the number of missing words to be guessed being dependent on sentence length. If a sentence is 50 words or less in length, the number of words to be guessed varies between one and ten. Given this, a score of '1' for a short sentence is not equivalent to a score of '1' for a longer sentence. To address this issue, results have been scaled to produce an adjusted score using the following formula:

$$\text{adjusted score} = \frac{(score + 1)}{gaps}$$

Adding 1 to the score ensures that '0' results are also scaled depending on the number of gaps in the sentence. Figure 18 compares adjustment of the score for a simple proportion (i.e. score/gaps) as compared to the formula above. A simple proportion does not produce a reasonable scaling, whereas the selected formula smoothly adjusts results by number of gaps. If the score is adjusted by a number greater than one, more extreme scaling is obtained. The optimal level of scaling may be different to that chosen, but in the absence of an external metric, the scaling to be chosen in our study is essentially an arbitrary choice. The adjustment however improves on raw scores or simple proportional adjustment.
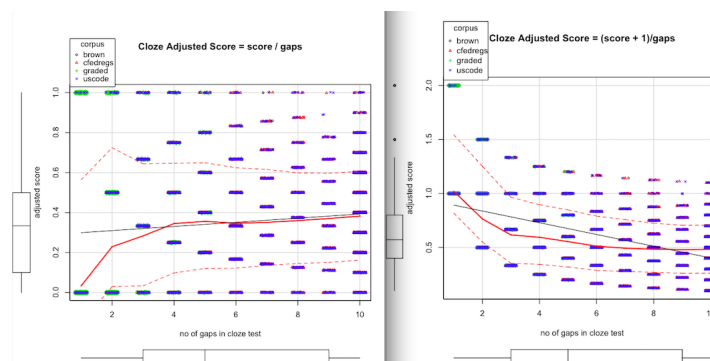


*Figure 18.* Comparison of cloze score adjustment schemes

An optimal scaling model merits further investigation but limitations of time made made this impractical.

The resulting ordering of sentences is different to the ordering established by the likert first principal component but is moderately correlated with it at -0.54. This level of correlation is similar to the correlation between the likert first principal component and the results discussed below from the semantic differentials (which are not affected by scaling issues).

Citizen Science for Citizen Access

Extending the principal components analysis described above to include the adjusted cloze results as an additional input variable produces a new first principal component, which correlates with the likert first principal component result at 0.98.

## 5.3. Semantic differential results

Semantic differentials can be used to derive a measure of user experience, which can also be broken down into a more nuanced set of characteristics. Semantic differential results were collected on responses to ten different semantic opposites. The semantic opposites were primarily chosen to capture user experience of using law, but also sought to explore the three dimensions which were identified by Osgood in his studies of semantic differentials: evaluative (good-bad); power (strong-weak): activity (active-inactive).

Six of the semantic differentials addressed concerns central to usability or user experience or readability of law. Two others addressed characteristics that users might associate with law: fairness-unfairness and attraction-repulsion. Both are evaluative, but evaluate law against notions of equity or emotional response to the content of the law. These two characteristics were chosen to explore whether individuals thoughts/feelings about the content of law affects their assessment of its readability. A list of the semantic differentials used is provided below. A summary term is provided in brackets and the semantic differential is asterisked if concerned with usability/user experience/readability characteristics. Although the original scale was between -3 and +3, the scale was adjusted to range between 1-7. Also, where necessary, scales were flipped so that a higher result means increasing strength in the characteristic. e.g. a readability score of 1 indicates less readability than a readability score of 7. This recoding was to assist in analysing and communicating results.

attractive-repellant (attractiveness)

clear-obscure (clarity)*

fair-unfair (fairness)

familiar-strange (familiarity)*

helpful-unhelpful (helpfulness)*

interesting-dull (interest)

severe-lenient (leniency)

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

readable-unreadable (readability)*

complex-simple (simplicity)*

usable-unusable (usability)*

Pearson's correlations were calculated for both raw semantic differential scores and mean scores for sentences. The level of correlation follows broadly the same pattern, with higher correlations found for the averaged scores (which is consistent with averaging out individual response variance). Table III shows correlations for the six semantic differentials associated with user experience and also includes correlation with the average adjusted cloze score and the likert first principal component. As will be evident, correlations across the table are moderate to high for most given characteristics, though the cloze average correlation was low for clarity, familiarity, helpfulness and readability and moderate with likert, readability and simplicity.

The pattern of correlation suggests that user experience characteristics including readability are not perfectly aligned although these characteristics have some degree of correlation. They also suggest that different testing methods will evoke different patterns of responses from users. Notably, the subjective measures (semantic differentials and likert tests) align to a greater degree than does the objective measure (cloze testing), although the likert and cloze results correlate moderately as between themselves. Among the semantic differentials we used, clarity, simplicity, readability and familiarity showed the highest correlation with the likert and cloze results. Also readability and clarity were highly correlated (0.83) and helpfulness and usability were also highly correlated (0.83). Simplicity had low to moderate correlation with helpfulness and usability (0.35 and 0.37, respectively). The differences in correlation indicate the more detailed description of user experience that semantic differentials can provide.

It is also of interest to examine distributions of responses for factors not associated with user experience, in this case looking at frequency of raw scores submitted by users. The diagonal on figure 19 shows this distribution. In most cases, users did not regard factors such as leniency or fairness as being relevant to assessing a test sentence. This may be compared to the quite different distribution for clarity, which did evoke mainly positive or negative assessments of clarity.

Citizen Science for Citizen Access

Table III. Pearson's R correlation for adjusted average cloze score, likert first principal component and six semantic differentials most related to user experience.

|  | clozemean | likertPC1 | clarity | familiarity | helpfulness | readability | simplicity | usability |
|---|---|---|---|---|---|---|---|---|
| clozemean | 1.00 | -0.55 | 0.34 | 0.33 | 0.13 | 0.46 | 0.53 | 0.13 |
| likertPC1 | -0.55 | 1.00 | -0.59 | -0.53 | -0.38 | -0.65 | -0.62 | -0.40 |
| clarity | 0.34 | -0.59 | 1.00 | 0.70 | 0.74 | 0.83 | 0.62 | 0.74 |
| familiarity | 0.32 | -0.53 | 0.70 | 1.00 | 0.62 | 0.69 | 0.51 | 0.65 |
| helpfulness | 0.13 | -0.38 | 0.74 | 0.62 | 1.00 | 0.64 | 0.35 | 0.83 |
| readability | 0.46 | -0.65 | 0.83 | 0.69 | 0.64 | 1.00 | 0.68 | 0.65 |
| simplicity | 0.53 | -0.62 | 0.62 | 0.51 | 0.35 | 0.68 | 1.00 | 0.37 |
| usability | 0.13 | -0.40 | 0.74 | 0.65 | 0.83 | 0.65 | 0.37 | 1.00 |
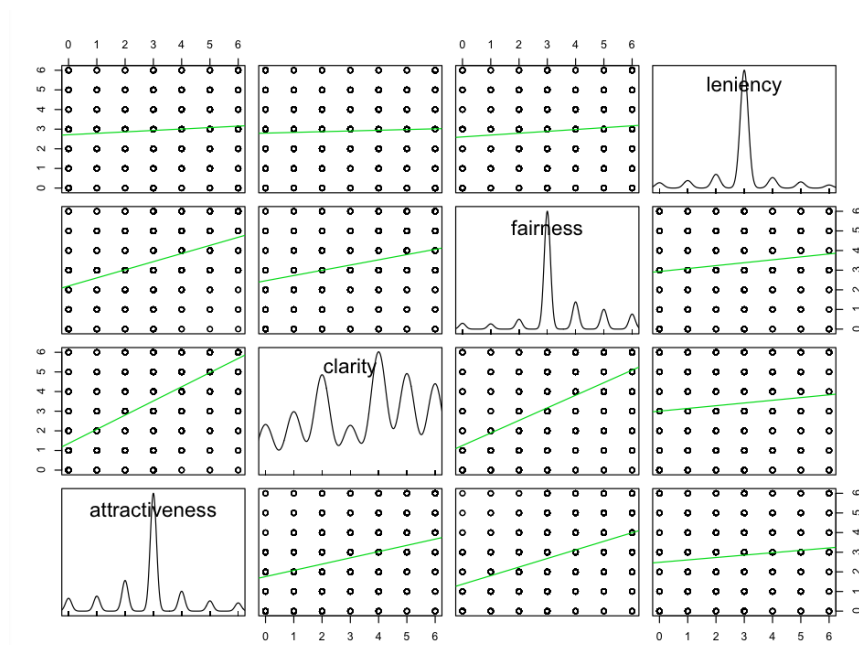


*Figure 19.* Density distribution of non-user experience characteristics

## 5.4. A Total Composite Readability Measure - Multivariate Analysis

In this section, we develop a total composite measure of readability. As with the individual measures, this measure can be developed by em-

ploying principal components analysis to extract the primary dimension of variance.

Table IV show the correlation between various composite measures of sentence difficulty. Clozemean1 is the mean adjusted cloze result described above. Semdiff4PC1 is the first principal component of semantic differential results using only the four semantic differentials which were most highly correlated with the cloze and likert results for this input: i.e. clarity, familiarity, readability and simplicity. Notably, these are semantically the most similar to the concept of easy/hard to read used in likert testing (usability and helpfulness being the other relevant terms which were less correlated). LikertPC1 is the first principal component of the likert results. CompositePC1 is the first principal component of the combined results from all tests. This last measure was derived from all 20 likert variables, the four semantic differential variables and the cloze mean adjusted score. The resulting composite measure is highly correlated with the principal component for semantic differentials and likert results, and moderately correlated with cloze mean results. The scree plot for this composite measure is also shown below (Figure 20), and as in the case of the Likert results, the first principal component is the sole component that satisfies the scree plot test.

Table IV. Correlation between different measures of 'readability'

|                | clozemean1 | compositePC1 | likertPC1 | semdiff4PC1 |
| -------------- | ---------- | ------------ | --------- | ----------- |
| clozemean1     | 1.00       | -0.64        | -0.55     | -0.47       |
| compositePC1   | -0.64      | 1.00         | **0.92**  | **0.90**    |
| likertPC1      | -0.55      | 0.92         | 1.00      | 0.68        |
| semdiff4PC1    | -0.47      | 0.90         | 0.68      | 1.00        |

The final composite measure can be used to generate an ordering of sentences which can be used to assign sentences to an 'easy' and 'hard' difficulty classifications. A 50% dividing line was used, thus 'easy' simply means the easiest half of the sentences, and 'hard' the hardest half. The assigned classifications can then be used as an input to machine learning.

It is worth commenting at this point that throughout the data, high variance was encountered in individual responses. How hard a user perceived or experienced a sentence to be, varied widely for both subjective and objective tests. The ability to derive reliable comparative measures of sentence difficulty therefore depends on being able to collect a sufficient number of assessments from users for each individual sentence.

Citizen Science for Citizen Access

**.PC**
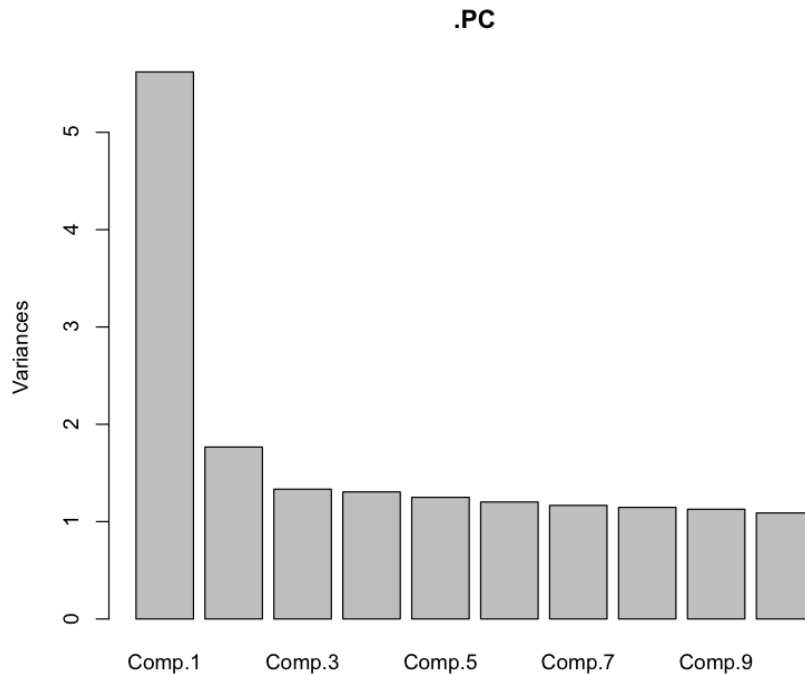


*Figure 20.* Composite principal components scree plot

Our results suggest that at least 15-20 separate user evaluations need to be collected for each test sentence.

## 6. Machine Learning

Collecting user evaluations of sentences is time consuming and difficult. It requires the availability of online infrastructure and access to audience. It calls on the time of users who are asked to provide evaluations. As this study has demonstrated, it is feasible to collect such user evaluations and this can provide valuable insight into the readability of legal language. Ideally however, we would wish to be able to predict the reading difficulty of a sentence without having to conduct surveys.

As has been noted, we divided the sentences into two equally sized "easy" or "hard" classes, depending on the sentence ranking according to the composite measures described in Section 5.4. To investigate

application of machine learning we extracted natural language features from the sentences including:

**(a)** sentence length;

**(b)** average word length;

**(c)** type to token ratio (i.e. ratio of unique words to total words);

**(d)** common readability metrics;

**(e)** proportion of verbal phrase chunks; and

**(f)** proportional distribution of different parts of speech.

Tests were carried out on two datasets: a dataset of all four corpora and a dataset of the sentences from just the two legal corpora. After a number of trials to investigate which machine learning algorithm achieved the highest level of accuracy for this task; a support vector machine (SVM) was chosen for the learning task.[15] All tests were validated using 10-fold cross validation.

## 6.1. RESULTS FOR FOUR CORPORA DATASET

An accuracy (F-measure) of was 72.7% was achieved. The overall result is less than the results reported at sentence level by Dell'Orletta et al which are discussed above in Section 2.6, who report an accuracy of 78% on the task of classifying sentences.

We also investigated the effect of removal (ablation) of particular features on prediction accuracy. In particular, we explored the contribution of traditional readability metrics to accuracy of machine learning.

### 6.1.1. *Effect of Ablation on Machine Learning Accuracy for Four Corpora Dataset*

For the whole dataset accuracy was reduced to 67.1% if only readability metrics were used as input. This was about the same accuracy as was achieved using just sentence length and average word length (67.3%). This is not surprising as readability metrics depend heavily on these

---

[15] That is, using the SMO package which is the support vector machine implementation in the Weka software. All reported results are for an SVM. The intuition behind an SVM is that (in a two dimensional case) the algorithm seeks to find the dividing line that maximises the distance of data points from the dividing line. In a case with many input features (which is usual for machine learning), the 'dividing line' is actually a hyperplane and each input feature is a dimension of a multidimensional space.

Citizen Science for Citizen Access

Table V. Precision, Recall and F-measure Support Vector Machine (SMO) on Four Corpora Dataset

| Class | Precision | Recall | F-Measure |
|---|---|---|---|
| easy | 0.737 | 0.705 | 0.72 |
| hard | 0.718 | 0.749 | 0.733 |
| Weighted Avg. | 0.727 | 0.727 | 0.727 |

two features. Using all features *except* readability metrics resulted in a prediction accuracy of 72.5%, i.e. virtually the same as including readability metrics with other as input features. Readability metrics can be concluded to be useless in predicting classification from the composite measure we used.

### 6.1.2. *Precision, Recall and F-measure Support Vector Machine (SMO) on Legal Corpora Dataset*

On purely legal sentences accuracy was 70.5%, i.e. a little less than for the four corpora dataset. Again this matches expectations as the graded sentences are virtually all in the easy dataset.

Table VI. Precision, Recall and F-measure Support Vector Machine (SMO) on Legal Corpora Dataset

| Class | Precision | Recall | F-Measure |
|---|---|---|---|
| easy | 0.691 | 0.571 | 0.625 |
| hard | 0.713 | 0.807 | 0.757 |
| Weighted Avg. | 0.703 | 0.705 | 0.7 |

### 6.1.3. *Effect of Ablation on Machine Learning Accuracy for Legal Corpora*

For the legal dataset accuracy was reduced to 60.2% if only readability metrics were used as input. Using all features except readability metrics resulted in an accuracy of 70.5%. For the legal dataset accuracy was 56% using just sentence length and average word length (i.e. machine learning essentially failed). Using just average word length, sentence length and type to token ratio achieved an accuracy of 66%. While using just phrase proportions and parts of speech proportions attained an accuracy of 67.8%

To further investigate the relationship between language difficulty and readability metrics we examined correlation between readability metrics and the composite difficulty measure. The SMOG index was most highly correlated at 0.33, which was about the same as correlation for sentence word length. This was however exceed by the type to token ratio at -0.42. In other words, it is more effective to count the ratio of unique words to total words as a measure of language difficulty of legal sentences than to rely on readability metrics at sentence level. Further, both for the four corpora dataset and the legal corpora dataset more accurate predictions can be obtained by a machine learnt model, than by using traditional readability metrics.

## 6.2. DISCUSSION OF MACHINE LEARNING RESULTS

The results of machine learning show the feasibility of improving accuracy of readability prediction over traditional readability metrics (i.e. 70.5 versus 60.2). This result is consistent with findings reported in the research literature. As far as we are aware this result has not been applied previously to legislative language (which is of course a unique form of English). We were able to show that accuracy can be increased on legal sentences by about 10% over use of traditional readability metrics alone. The overall level of accuracy of 70.5% is lower than that reported by other researchers on sentence level classification. It may be possible to increase the level of accuracy by extracting more complex natural language features. Also, it is likely that increasing the amount of data on which learning can be carried out would also increase accuracy. The rate at which we could generate human labelled data was as limiting factor in our study.

## 7. Conclusions

## 7.1. APPLYING CITIZEN SCIENCE TO READABILITY OF LEGISLATIVE TEXTS

The research reported in this paper demonstrates the feasibility of using citizen science (in the form of online crowdsourced data collection) to create a corpus of labelled data for input to machine learning for predicting the readability of legal rules. It is possible to rank a given set of legal sentences by reading difficulty using responses submitted by users. However, the time required to collect the necessary data is non-trivial, even on a large sites such as LII. Data in our study was collected over a three month period. Even after three months, the sentences tested represents a tiny proportion of the legal language

Citizen Science for Citizen Access

found in the United States Code and Code of Federal Regulations. Extending this research to a larger dataset would potentially require orders of magnitude longer. This is far from a fatal barrier, and does give insight as to what may be required to collect sufficient data to better predict the readability of legal language. However, considering the time horizons that are sometimes necessary for research in other fields (for example longitudinal health studies, or multiyear collection of astronomical data), it is well within the bounds of realistic research. An observation that bears on this conclusion is that participant responses were maintained throughout the period of collection, suggesting that it is feasible to collect data over long periods, without the rate at which data is collected reducing over time. A novel (or at least unusual) aspect of our citizen science project was that citizen scientists assisted in collecting information about themselves as well as about the 'data'.

## 7.2. Demographic insights

The demographic results described above are also of interest. We can begin to reach conclusions about who reads the law and why they do so. Legal professionals (including law students) were a minority of those who participated as citizen scientists in our research. Determining whether this is true of the online audience for legislation generally, requires further research, including on other online sites. Nonetheless, a substantial audience for online legislative materials are non-lawyers. Non-lawyers find legislative materials harder to access that the legally trained. If we wish to communicate effectively with this substantial audience, we need to re-examine how the law is written. Women, those not having tertiary education and those for whom Spanish is the primary language were under-represented among participants. Gender, education and language aspects of access merit further investigation. Our study is consistent with the findings of other researchers that legislative language is harder for those without legal training. It also suggests that legislative language is hard for all audiences, including the legally trained.

## 7.3. Machine learning

Our work on machine learning reports initial application of machine learning to the readability of legislative materials. We have demonstrated that traditional readability metrics can be improved on, for high resolution (i.e. sentence level) automatic classification of legal sentences into a binary easy vs. hard classes. The level of accuracy attained is moderate and would require further improvement to provide a reasonably usable automated detection system. We are planning to publish a

second paper that extends the machine learning results reported here, including investigating how accuracy may be increased by increasing the number of input features and further exploring whether increasing the available data may assist in improving accuracy.

## 7.4. Methods of Measuring Readability Using Crowdsourced Data

Part of this paper is devoted to describing the methods we used to collect crowdsourced assessments of readability. A considerable portion also describes the methods used to convert crowdsourced assessments into measures of sentence readability. In part this goes to reproducibility of the research. However, it also seemed useful to us to describe our methods at some length, as there is no 'standard' method for carrying out crowdsourced readability research. It is useful for the research literature to provide descriptions of this kind. Further, we have no doubt that the methods reported in this study can be improved on.

Of the three tests that we used, likert tests proved to be the most effective in attracting participation and in ensuring data was usable. Semantic differentials provided a more nuanced characterisation of the sentences being tested. However, the rate of response for semantic differentials was much lower than likert tests and the occurrence of unreliable data much higher. Cloze tests, unlike the previous two tests, had the advantage of providing an objective measure and proved to be the only useful test for distinguishing readability for different demographic groups. However, analysis of cloze results was complex and like semantic differentials they attracted a lower response rate. Again issues of data reliability reduced the usable data. Also given wide differences in sentence length, cloze tests were not ideally suited to sentence level assessment. A limiting factor that emerged in the study, for all three methods, was the rate at which assessments could be collected. Methods which reduce the necessity for replication may significantly increase the rate of data collection.

## 7.5. How readers read the law online

We may be all equal before the law but the law is not equally of interest to its readers. In fact, the frequency with which a particular piece of law is read follows a power law distribution. This is an important insight. If we are concerned with improving reader experience, attention to that part of the law which is most read, provides exponentially greater return and requires fractional effort as compared to seeking to improve the law book as a whole. Further, if law is not being read, we may ask the question: how important is it for that law to remain in the

Citizen Science for Citizen Access

law book? At least, in terms of organising legal documents, those parts which are most read, might be usefully reorganised in ways that make them more accessible to readers.

7.6. SOME BROADER IMPLICATIONS

The data repositories and online publishing platforms which sites such as Cornell LII maintain, can perhaps be thought of as potentially playing enhanced roles in improving access to law. Such sites have achieved access to law in terms of ensuring that citizens are able to find and access law online. The fact of availability does not, however, necessarily equate to "access" in all its senses. Addressing the readability of legislation by applying online technologies is a natural extension of the work already carried out by the Free Access to Law Movement.

Online legal publishing platforms are also potentially sites for the on-going collection of data which illuminates how users interact online with legal language. They are not simply collections of text or collections of data, they are a focus of a dynamic and ongoing interaction between human beings and the laws that govern them. We can perhaps trace the outlines of a paradigm in which the publication of law online - already moving from being conceived as static document to data repositories - is reconceptualised even further as an online platform capturing a multiplicity of points of human-legal interaction with the potential to tell us a great deal about the social dimensions of law. Or, in other words, online law is part of a social network in which both human beings and legal rules (communicated by other human agents) are nodes. The insights that we may derive from a study of these interactions could over time be applied to improve legal language – addressing an as yet unmet dimension of making law accessible to all who would like to have that access. To extrapolate from the words of the UK Parliamentary Counsel: when citizens find the law, they should able to read it. Other applications outside the readability field may also exist.

## 8. Future Work

We are interested in extending the work reported here into the following areas of research:

**(a)** extending citizen scientist participation in other aspects of readability research (for example project design);

**(b)** investigating other means of collecting readability assessments of legal language online, for example A-B testing, a simplified form

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

of likert or approaches that calibrate between different testing approaches;

**(c)** further investigating the demographic aspects of access to law online, particularly gender, education and language; and

**(d)** extending the preliminary machine learning results reported in this paper.

## References

Eloise Abrahams. Efficacy of plain language drafting in labour legislation. Master's thesis, 2003.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, 2010.

Gregory Asmolov. Crowdsourcing as an activity system: Online platforms as mediating artifacts. In *Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*, 2014.

Robert W Benson. End of legalese: The game is over. *NYU Rev. L. & Soc. Change*, 13:519, 1984.

J. Bentham. Nomography, or the art of inditing laws. *The Works of Jeremy Bentham (ed. J. Bowring)*, 3:231 et seq, 1843.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python.* O'Reilly Media Inc, 2009.

J.R. Bormuth. *Cloze readability procedure.* University of California Los Angeles, 1967.

F. Bowers. Victorian reforms in legislative drafting. *Tijdschrift voor Rechtsgeschiedenis*, 48:329, 1980.

Daren C Brabham. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, 14(1):75–90, 2008.

Dennis L Clason and Thomas J Dormody. Analyzing data measured by individual likert-type items. *Journal of Agricultural Education*, 35:4, 1994.

Citizen Science for Citizen Access

Kevyn Collins-Thompson. Computational assessment of text readability: A survey of past, present, and future research. *working draft*, July 2014.

Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

Anna B Costello and Jason W Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7):2, 2005.

Mick P Couper, Roger Tourangeau, Frederick G Conrad, and Eleanor Singer. Evaluating the effectiveness of visual analog scales a web experiment. *Social Science Computer Review*, 24(2):227–245, 2006.

Michael Curtotti and Eric McCreath. Enhancing the visualization of law. In *Law via the Internet Twentieth Anniversary Conference, Cornell University*, 2012.

Michael Curtotti and Eric McCreath. A right to access implies a right to know: An open online platform for research on the readability of law. *Journal of Open Access to Law*, 1(1), 2013.

O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 2013.

Joost CF de Winter and Dimitra Dodou. Five-point likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11):1–12, 2010.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83. Association for Computational Linguistics, 2011.

Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4): 86–96, 2011.

W.H. DuBay. The principles of readability. *Impact Information*, pages 1–76, 2004.

M. Evans and R.I. Jack. *Sources of English Legal and Constitutional History*. Butterworths, 1984.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

Carolina Ferrari, Tomas Garcia & Short. Legibility and readability on the world wide web, 2002. URL `http://bigital.com/english/files/2008/04/web_legibility_readability.pdf`.

Frank J Floyd and Keith F Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286, 1995.

John Fox. The R Commander: A basic statistics graphical user interface to R. *Journal of Statistical Software*, 14(9):1–42, 2005. URL `http://www.jstatsoft.org/v14/i09`.

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, 1964. URL `http://www.hit.uib.no/icame/brown/bcm.html`.

Michael S Friman. Plain english statutes-long overdue or underdone. *Loy. Consumer L. Rep.*, 7:103, 1994.

Edward Fry. A readability formula for short passages. *Journal of Reading*, 33(8):594–597, May 1990.

Ron Garland. A comparison of three forms of the semantic differential. *Marketing Bulletin*, 1(1):19–24, 1990.

GLPi and V. Smolenka. A Report on the Results of Usability Testing Research on Plain Language Draft Sections of the Employment Insurance Act, 2000. URL `http://www.davidberman.com/wp-content/uploads/glpi-english.pdf`.

Elizabeth M Grieco, Yesenia D Acosta, G Patricia de la Cruz, Christine Gambino, Thomas Gryn, Luke J Larsen, Edward N Trevelyan, and Nathan P Walters. The foreign-born population in the United States: 2010. *American Community Survey Reports*, 19:1–22, 2012.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software. *SIGKDD Explorations*, 11(1), 2009.

Citizen Science for Citizen Access

Richard H Hall and Patrick Hanna. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23 (3):183–195, 2004.

Eric Hand. Citizen science: People power. *Nature International Weekly Journal of Science*, 466:685–687, 2010.

Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Published online, 2003.

J. Harrison and M. McLaren. A plain language study: Do New Zealand consumers get a "fair go" with regard to accessible consumer legislation. *Issues in Writing*, 9:139–184, 1999.

Richard M Heiberger and Naomi B Robbins. Design of diverging stacked bar charts for likert scales and other applications. *Journal of Statistical Software submitted*, pages 1–36, 2013.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics, 2008.

P. Heydari and A.M. Riazi. Readability of texts: Human evaluation versus computer index. *Mediterranean Journal of Social Sciences*, 3 (1):177–190, 2012.

Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

Simon James and Ian Wallschutzky. Tax law improvement in Australia and the UK: the need for a strategy for simplification. *Fiscal Studies*, 18(4):445–460, 1997.

Frances Johnson. Using semantic differentials for an evaluative view of the search engine as an interactive system. In *EuroHCIR*, pages 7–10, 2012.

Sasikiran Kandula and Qing Zeng-Treitler. Creating a gold standard for the readability measurement of health texts. In *AMIA Annual Symposium Proceedings*, volume 2008, page 353. American Medical Informatics Association, 2008.

Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International*

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

*Conference on Web Search and Data Mining*, volume WSDM '09, February 9-12, 2009, Barcelona, Spain., pages 202–211. ACM, 2009.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554. Association for Computational Linguistics, 2010.

David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. Text simplification tools: Using machine learning to discover features that identify difficult text. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 2616–2625. IEEE, 2014.

J. Kimble. Answering the critics of plain language. *Scribes J. Leg. Writing*, 5:51, 1994.

G.R. Klare. Readable computer documentation. *ACM Journal of Computer Documentation (JCD)*, 24(3):148–168, 2000.

S. Krongold. Writing laws: Making them easier to understand. *Ottawa L. Rev.*, 24:495, 1992.

Bruce Lewenstein. What does citizen science accomplish. 2004.

D. Melham. Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs, 1993.

David Mellinkoff. *The Language of the Law*. Little, Brown and Company, 1963.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics, 2010.

Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

Citizen Science for Citizen Access

PCO NZ. Presentation of New Zealand Statute Law: Issues Paper 2. Technical Report 2, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office, 2007.

PCO NZ. Presentation of New Zealand Statute Law. Technical Report 104, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office, 2008.

Law Reform Commission of Victoria. Access to the Law - the structure and format of legislation. Technical Report 33, Law Reform Commission of Victoria, 1990.

OLR. Inland Revenue Evaluation of the Capital Allowances Act 2001 rewrite, Opinion Leader Research. Technical report, UK Inland Revenue, 2003.

OPC-Australia. Plain English. Technical report, Australian Commonwealth Office of Parliamentary Counsel, 2003.

OPC-UK. When Laws Become Too Complex: A Review into the Causes of Complex Legislation. Technical report, United Kingdom Office of Parliamentary Counsel, 2013.

N. Pettigrew, S. Hall, and D. Craig. The Income Tax (Earnings and Pensions) Act - Post-Implementation Review, Final Report MORI, 2006.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics, 2008.

Marta Poblet, Esteban García-Cuesta, and Pompeu Casanovas. IT enabled crowds: Leveraging the geomobile revolution for disaster management. In *Proceedings of the Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*, pages 16–23, 2014.

R-Core-Team et al. R: A language and environment for statistical computing. 2012.

D Renton. The preparation of legislation - report of a committee appointed by the Lord President of Council. Technical report, Council of the UK Government, 1975.

Curtotti, Weibel, McCreath, Ceynowa, Frug, Bruce

Camille L Ryan. Language use in the United States: 2011 American community survey reports. *Washington, DC: US Census Bureau*, 2013.

Camille L Ryan and Julie Siebens. Educational attainment in the United States: 2009. *Washington, DC: US Census Bureau*, 2012.

Adrian Sawyer. Enhancing compliance through improved readability: Evidence from New Zealand's rewrite "experiment". *Recent Research on Tax Administration and Compliance*, 2010.

Sarah E Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2005.

Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM, 2001.

Johan Sjöholm. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. PhD thesis, Linköping, 2012.

D. Smith and G. Richardson. The readability of Australia's taxation laws and supplementary materials: an empirical investigation. *Fiscal Studies*, 20(3):321–349, 1999.

Edwin Tanner. Seventeen years on: Is Victorian legislation less grammatically complicated. *Monash UL Rev.*, 28:403, 2002.

Carol Tullo. Solving the challenge of the 21st century statute book. In *Law via the Internet Conference*, 2013. URL `http://www.jerseylvi2013.org/presentation/solving-the-challenge-of-the-21st-century-statute-book/`.

Arnold Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 521–530. ACM, 2010.

G. Wagner. Interpreting cloze scores in the assessment of text readability and reading comprehension, 1986.

Citizen Science for Citizen Access

Andrea Wiggins and Kevin Crowston. From conservation to crowd-sourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

B. Woods, G. Moscardo, T. Greenwood, et al. A critical review of readability and comprehensibility tests. *Journal of Tourism Studies*, 9(2):49–61, 1998.

## Acknowledgements

# INTERDISCIPLINARY COOPERATION IN LEGAL DESIGN AND COMMUNICATION

## Michael Curtotti[1], Helena Haapio[2], Stefania Passera[3]

[1]PhD Researcher, Australian National University, Research School of Computer Science, Legal Counsel, ANU Students' Assocation & ANU Postgraduate and Research Students' Association
Building 17a Student Facilities Building, Union Court, Australian National University, ACTON, ACT 0200, AU
michael.curtotti@anu.edu.au
[2]Postdoctoral Researcher, University of Vaasa / International Contract Counsel, Lexpert Ltd
Pohjoisranta 20, 00170 Helsinki, FI
helena.haapio@lexpert.com; http://www.lexpert.com
[3]PhD Researcher, Aalto University School of Science, Department of Industrial Management and Engineering
Betonimiehenkuja 5C, 02150, Espoo, FI
stefania.passera@aalto.fi; http://legaldesignjam.com

*Keywords:* ***legal design, legal visualization, legal communication, multidisciplinary collaboration, contract visualization, cross-professional communication***

*Abstract:* *The last two decades have seen law emerge online. This development has engaged computer scientists and web designers in communicating law. Recently, serious work has begun on visualizing contract clauses, generating cooperation between designers, computer scientists, business people, lawyers and others. New insights arise from such cross disciplinary collaborations. Each discipline provides theoretical insights as to how legal design and communication might be approached. More profoundly each has the potential to recast relationships – what does it mean for the 'power' of law makers to be exercised in the context of such paradigms? How do such insights enable us to reconsider the role of lawyers: the traditional custodians of legal rules? We examine these questions from a theoretical viewpoint, and reflect on our own cross-disciplinary collaboration in the creation of a proof-of-concept tool for automation of contract visualization.*

## 1. Introduction

Historically, legal rules have been the exclusive preserve of lawyers and the legally literate. We live in a time where this is no longer true. The online environment has necessarily involved the engagement of computer scientists and software designers in legal communication. More, it has made law accessible to a much broader audience. By querying legal communication in the context of professional affiliation we place it within a social context which may affect its form, content and manifestation. A focus on a computational context leads us to look at law from the viewpoint of software developers and to see laws as data. A business context, in turn, expands our view from data to information and knowledge and how these are designed and communicated. Visualization calls on a design paradigm, and on understanding the skills, needs, and experiences of users in order to craft information into usable visual artefacts. This contextual exploration also allows us to re-examine the nature of legal rules. The catalyst that led to this paper was our previous collaboration to create a prototype tool to automate the visualization of selected contract clauses used in business to business contracts. This work involved a multidisciplinary collaboration bringing together design, legal

and computer science paradigms. [Passera et al. 2014][1]

## 2. Thinking Like Designers

In his book *The Design of Everyday Things* Don Norman observes that *"[a]ll artificial things are designed"* [Norman 2013, 20]. He states: *"[Design] focus[es] on the interplay between technology and people to ensure that the products actually fulfill human needs while being understandable and usable ... not only must the requirements of engineering, manufacturing, and ergonomics be satisfied, but attention must be paid to the entire experience, which means the aesthetics of form and the quality of interaction."* [Norman 2013, 20]

This description evokes Patrick Jordan's hierarchy of product user needs, which traverse *functionality, usability* and, at the apex, *user experience.* [Norman 2013, 13; Haapio 2013] Human centred design *"... puts human needs, capabilities, and behavior first, then designs to accommodate those needs, capabilities, and ways of behaving."* [Norman 2013, 24] Two further concepts that Norman explores are relevant to the task of addressing user needs: *affordances* and *signifiers.* The latter concept focusses on communication as part of good design. *"Affordances represent the possibilities ... for how an agent (a person, animal, or machine) can interact with something. ... Signifiers are signals. Some signifiers are signs, labels, and drawings ... Some signifiers are simply the perceived affordances, such as the handle of a door ..."* [Norman 2013, 28, 31–33] Human centeredness and design addressing a hierarchy of needs and facilities such as affordances and signifiers are not necessarily familiar in the traditional design of legal artefacts.

## 3. Thinking Like Computer Scientists and Software Engineers

Early in their studies, most students of computer science will be required to write computer programs. Further they will be taught a number of conceptual paradigms for expressing programs. Among the many alternatives are procedural, functional and object oriented. [Thompson 1999; Horstmann 2006; Chun 2007] Each paradigm provides a different way of solving the same problem. Procedural (or imperative) programming uses "declarations" (definitions) and "statements" (commands). The way they are expressed is similar to the way legal rules are expressed. Naturally so, as the paradigm imagines the computer as a digital agent and sets out the commands the agent must carry out. In functional programming, program statements are structured as interlinked mathematical functions. In object oriented programming, virtual objects having properties and capacities interact with each other within the program. Although the design is quite different in each case, the output may be identical. As the computer speaks its own underlying 'machine language' it is indifferent to how the source code is written. Thus, as well as communicating with the computer, computer scientists are communicating with each other. The communication is multidirectional and multipurpose. Programming paradigms assist software engineers to more effectively create, reproduce, maintain and collaborate in creating potentially vast software artefacts.

To lawyers, laws are legal rules[2] and documents. To engineers or computer scientists laws may be looked at as data. The software product that surrounds that data can take a multiplicity of forms. It is determined by those commissioning the software. We can see how engineering influences the external manifestation of law in examples drawn from its online publication.

Law may be represented online as a pdf document. Each pdf corresponds to the traditional *physical*

---

[1] The prototype can be accessed at http://cs.anu.edu.au/people/Michael.Curtotti/visualcontracting/.
[2] The terms "laws" and "legal rules" here are labels for 'law' found in contracts, legislation and regulations. The use of the term has not intended to say anything about the nature of law, although we discuss this in section 6 below.

publication of a 'compilation' which includes all amendments to a particular date.[3] The early 1990's saw laws presented as hyperlinked documents – with each section represented as a separate hyperlinked document. Each section can be provided with multiple links – such as to preceding and following sections, to defined terms, and to relevant case law and journal articles.[4] Recently the publishers of UK legislation have explicitly regarded legislation as data. In addition to online readable versions of the law, the law is made available as data in its raw xml format. The native form of the law is a hierarchical data structure and as such, it can be used in potentially infinite forms of software. [Curtotti & McCreath 2012]

In the creation of a software product, software engineers are also concerned with the viewpoint of the user: seeking to satisfy customer expectations. Here the software developer, within the functional constraints of the software, thinks much like a designer.

An example is afforded by Australia's migration law. The law itself is fairly described as a forbidding and impenetrable morass. This mass of legal rules is converted on the government's website into a usable user interface. The primary concerns of users to potentially visit, live, study or work in Australia are prioritised. Visa options that address these needs are prominently displayed in accessible thematic lists. Individual visa pages provide concise easily navigated tabs providing an overview of the visa, eligibility criteria, how to apply and the holder's rights and obligations. An online application process is also provided. While under the hood, legal "rules" (the legislative "code") govern and define the process, the legal details are re-organised, with irrelevant and less relevant information hidden, significantly improving user experience as compared to attempting to use the underlying law which it represents.[5]

An example of a *visual user interface* facilitating the creation and use of legal rules is provided by the Creative Commons copyright licensing system. [Haapio 2013, 73] Four, now widely recognised symbols, are used to capture the intent of the licensor.[6] These symbols are easy to learn but convey key aspects of Creative Commons licences.

Both of the above cases are examples of *abstraction*. In software engineering terms this is "*suppressing or ignoring some properties of objects, events, or situations in favor of others*". [Fox 2006, 7] The non-essentials are hidden or *encapsulated*, while essential features are made manifest. [Horstmann 2006, 67] Creative Commons is also an example of the use of legal rules to enable rather than to regulate and control. It empowers an entire ecosystem of use and re-use of copyright content.[7] As of 2009, it was estimated that 300 million works had been licensed under the Creative Commons system. [Kapitze 2009, 104]

A number of insights may be drawn from the foregoing discussion. The traditional form (paradigm) in which legal rules are expressed is not essential to them. Other forms of expression are possible. Legal rules may be thought of as internal aspects of a broader "legal/policy product". Such external characteristics can be used to address usability and user experience. Legal rules are also data. Enabling a full application of computational technologies requires that they be made available as data. An important tool in the task of legal communication is *selection of information*. Information hiding (abstraction and encapsulation) can aid communication.

## 4. Thinking Like Lawyers and Business Managers

It is striking that when contract law is taught to law students, engagement is avoided with the con-

---

[3] For example, Australian Capital Territory legislation website http://www.legislation.act.gov.au.
[4] For example, the United States Code published at http://lii.org is in this form.
[5] Australian Department of Immigration and Border Control web pages http://www.immi.gov.au.
[6] Creative Commons. About the Licences https://creativecommons.org/licenses/.
[7] See discussion below on the implications of design approaches for the theory of law and power.

tracts themselves. Rather contract law is largely taught as legal theory: when contracts arise; the rules for their interpretation; and the consequences of breach or termination. This can be seen in descriptions of law school curricula and in the contents pages of contract law textbooks.[8] The drafting of contracts themselves is not a focus of study, and contract clauses are seldom seen as part of "law" in legal education. This contrasts with how computer science students are taught about programs – as explored above. Some parallels can be drawn. As a software engineer seeks to ensure that a program deliver on requirements (that it functions correctly), a lawyer seeks to ensure that a contract is legally binding. However much of the law student's attention is directed to the legal consequences of contract failure. Further, legal scholarship around contracts has traditionally focussed on contract law in the context of litigation rather than on the correct operation of contracts within a business context. [Haapio 2013, 2, 6; Pohjonen 2009]

The creation of legislation (the public parallel to contract drafting) is the responsibility of a very small group of professional drafters (usually within government). Those involved in legislative drafting often state that it takes seven to eight years of practice to develop the skills of legislative drafting. The audiences include both professionals and lay users of the law. Yet when drafting the law, beyond functionality, the primary concerns are legislative intent and judicial interpretation: thus two audiences are primarily thought of: parliamentarians (with the goal that they will vote for the law) and judges (with the goal that their interpretations will give effect to the government's policy). *Within these constraints*, the drafter seeks clarity to avoid unnecessary litigation and cost.[9]

Of course lawyers, like other professions, are diverse. Some business lawyers are influenced by the views of business managers and see the goals of contracts differently from judges and litigation lawyers. For business, achieving the business objectives and succeeding in implementation are the goals, winning or resolving legal disputes is secondary. Many disputes are preventable through better contract design and communication. [Siedel & Haapio 2010; Haapio 2013; Haapio 2006] Contract law, education and research have concentrated on what *courts* have done, in hindsight, ex post. Businesses, again, seek successful transactions, and what *people and businesses* can and should do – foresight, ex ante. The typical law school education reinforces the notion that litigation is at the core of lawyering. Students spend a lot of their time reading about case law. Most contract law books are full of examples of failures; contracts that have become embroiled in a dispute or litigation. Traditional law is mostly reactive, and not many lawyers have questioned the habit of looking at precedents and the past, or of focusing on failures. [Pohjonen 2009; Haapio 2006] There seems to be a major gap between *academic law* and *law in action* and also between *contract law* and *contract practice*. [Mitchell 2013] These gaps need to be bridged.

## 5. Crossing Disciplinary Boundaries

The genesis of this paper was our own experience working in a cross disciplinary context to develop prototype implementations for automating the visualization of contract clauses. In undertaking this collaboration one of us worked primarily (though not exclusively) as software developer, one primarily as visual designer and one primarily as business lawyer.

For the software developer, it became evident that the *form/language* in which a clause is expressed is peripheral. A practical way of solving the problem of automating the visual representation of a clause was to identify the underlying variables (data) essential to a particular clause. These variables controlled the visual representation. It also became evident that certain types of clauses lent themselves more readily to automated visual representation. Those involving time or numerical quantities were obvious candidates. Clauses lacking such characteristics presented barriers to auto-

---

[8] E.g. 702011 Contracts Course University of Technology Sydney http://handbook.uts.edu.au/subjects/70211.html.
[9] Steven Laws cited in *Stefanou & Xanthaki* 2013, 20, 24–25.

mation. Like the case of Creative Commons, a visual language would need to be consciously created to effect automated visualization.

For the visual designer, in addition to identifying information characteristics that can be mapped visually, the question was how to visualize in ways that are understandable, engaging and useful for the intended users. Gestalt psychology principles,[10] for instance, are fundamental for the design of explanatory diagrams. Information needs to be structured in visual hierarchies that support unambiguous and fast understanding. In the case of automation, designers must understand through user studies and contextual inquiries how people work with contracts. In addition to their information needs and the design of the visual output, the whole interface and its functioning must be considered. Designers seek to address meaningfully the needs of users, and are not content to find a way to translate data or clauses into diagrams. How can such new tools work seamlessly with existing drafting tools? How can the interfaces be designed to be easy to use and learn, providing a feeling of control and trustworthiness to users? In what ways can visualization and automation *really* bring substantial benefits to users, i.e. is it about enhancing communication between parties, auditing one's understanding of clauses through visual means or exploring and comparing different ways in which a certain provision can be arranged? The answers to these questions come only by researching the users' reality and continuously validating possible solutions through prototyping, and usability and user experience studies.

For the business lawyer, legal and managerial requirements came to the fore. The prototype should be easy to use for managers and lawyers so as to generate text and images that are legally sound. In addition, it needed to support managers and lawyers in informed decision-making at two stages: 1) when the contract is planned and 2) at the contract implementation stage. At both stages, clarity as to the parties' rights and obligations is needed, and at neither stage should unintended liabilities or remedies arise from the implied, "invisible" terms which operate by law. While the contract clauses selected for our prototype were such that we did not need to address additional (as opposed to exclusive) remedies, we learnt much about these remedies and their interpretation under different legal systems when preparing the presentation of our paper at IRIS 2014.[11]

For the entire team it became clear that computer-assisted visualization is a powerful tool that may deal with data, information or knowledge, as well as with the needs and aspirations of different users who work with contracts. We were not dealing with the exploration and visualization of large volumes of data (information visualization). Instead, we were supporting the creation, application and communication of knowledge and insights. [Eppler 2004; Eppler & Burkhard 2004] For our experiment, knowledge visualization offered strategies, tools and methods to make contract-related knowledge accessible and visible and to improve processes through which knowledge can be identified, assessed, shared, discussed, applied and managed. [Haapio 2013, 13]

## 6. Rethinking the Nature of Legal Rules and the Role of Lawyers

*What is Law?* It is a question that has been the subject of extensive attention by legal theorists. But how do cross disciplinary insights affect our answer to this question? When legal thinkers have considered what they mean by *law – law as rule –* is often central. Three prominent theories are offered as illustration.

Legal positivism, particularly the command theory of law, holds that law can best be understood as general commands communicated by a recognised sovereign power which is habitually obeyed and

---

[10] Gestalt principles describe how the mind organizes perceptual scenes and discriminates visual stimuli, e.g. between parts and whole, figure and background. See: http://www.scholarpedia.org/article/Gestalt_principles.

[11] The presentation can be accessed at http://www.mindspace.fi/wp-content/uploads/2014/02/IRIS_passera_haapio_curtotti.pdf. For exclusive (only, sole) as opposed to additional remedies, see under Insights, especially slide 33.

which can punish disobedience. This view explicitly seeks to distinguish such rules from phenomena considered to be non-law: religious law, ethical precepts, customary usages, and social etiquette. [Harris 1980, 26; Bix 1999, 34] Natural law theory by contrast holds that certain universal legal norms exist in the abstract: norms which no human power can abrogate and which are inherently known by all human beings. [Harris 1980, 7–8] The most pervasive modern expression of natural law is human rights law, which although expressed in positive law instruments (treaties and national laws), also appeals to pre-existent universals.[12] Critical legal theory holds that behind a veneer of legitimate authority, laws are rules imposed by the powerful on the weak. The task of critical legal theory is to expose these political and social realities. Thus critical race theory understands that racism has profoundly shaped the legal system. Similarly feminist critical theory views law as patriarchal and oppressive to women. [Bix 1999, 203 et seq; Patterson 1996 et seq] In one way or another the concept of *rule* is at the centre of such theories.

From a design viewpoint, the *rules* or *constraints* of law are not central. First and foremost laws are designed artefacts. They are designed optimally or poorly. What is important is that they are designed in a way that is functional, usable and provides a positive user experience. The "rules" are not central. Key is that each user of the traffic system be empowered to effectively and safely navigate it. The system is provided with *affordances* and *signifiers* such as traffic lights and pedestrian crossings. Good design would also suggest that the design of traffic laws takes account of the characteristics of all users, not just those who drive luxury cars. Those who use legal rules, rather than those who make them, are at the apex. Legal rules are thus an enabling framework. The emergent characteristic of human empowerment is thus central to a *design* view of the law.[13] Somewhat surprisingly, a design paradigm suggests a democratising theory of law. Law making at its is best empowers citizens. How distant from a command theory of law![14]

Although space does not permit further exploration here, computer science takes us in entirely new directions in thinking about the nature of law. Law as data, law as functional core, and separation of legal form from legal function. Both paradigms however allow us to consider the role of lawyers. If law, taking a design view, is an instrument of empowerment rather than of control, what is for example the lawyer's role in contract design? The shift of the lawyer from a regulating role to an empowering role emerged clearly in our experiment with automating contract visualization. [Passera et al. 2014] Legal expertise is not required solely to create legally binding commitments or to respond to legal challenges one contract at a time, but rather to orchestrate in advance what to put in the end user's "toolbox", and how these tools should behave, look, and feel when used in real-life situations. It takes great knowledge to create a tool that can adapt to different situations without "breaking". In the case of a tool generating textual and visual versions of a clause, one needs to know all possible alternatives and instances of that clause, its meaning, use and role in a real context. Also discrimination is required between information which can be modeled in advance in a tool and elements needing human intervention. The lawyer-designer becomes principally concerned with *making contracts work for clients* and *empowering the clients to achieve their goals* [Haapio 2012; Pohjonen 2009; Pohjonen & Visuri 2008], rather than crafting legally enforceable obligations or minimizing legal risks. We see a much better fit with the business manager's perspective. Certainly some control of legal risk is necessary, but from a design viewpoint it is far from sufficient. Insights drawn from software engineering might shape what lawyers do or at least how they are trained. The current paucity of legal training devoted to addressing the kind of tasks in which software engineers

---

[12] E.g. Martin Luther King Jr. explained the US Declaration of Independence in natural law terms [King 1965].

[13] Emergence (e.g flocking (*Reynold's boids)*; or reproduction (*Conway's Game of Life*) are noted by computer scientists as potential properties of simple agent level rules. Emergence also occurs in biology [Bedau 1996, 12].

[14] Interestingly this view corresponds to one thread in feminist conceptions of power which distinguishes the power "to" from the power "over". "*The masculine 'power over' construct ... connotes a 'command-and-control' ... 'power to' envisions a more egalitarian and empowering type of leadership.*" [Reingold 1996].

are rigorously schooled, is unlikely to be best preparing lawyers to serve the needs of their clients. In our view, legal education would benefit from drawing on this disciplinary example.

## 7. Conclusion

Societal change is opening the law to new and diverse forms of exploration. New communication technologies are transforming the practice, theory, making and teaching of law. Experience and research indicate that law has entered an age where its design and communication must change. We need more user-friendly interfaces to law. The implications of the bodies of thought which have driven revolutions in design and in information technology are yet to be extensively explored in connection with the form, nature and content of legal rules. We have explored some aspects of how law is being influenced by these changes, highlighting relevant conceptual frameworks, and how they have already affected the law. Further, we have addressed our own cross-disciplinary collaboration. At a number of points in our paper we have also suggested directions or highlighted points which warrant further exploration. Visualization offers a promising way to facilitate effective cross-professional communication and collaboration: it helps improve law's usability and user experience. Continued cross-disciplinary research is needed to recognize the opportunities and challenges and benefit from these bodies of knowledge.

## 8. References

*Bedau, Mark A.* (1996). Philosophical Aspects of Artificial life. http://pdf.aminer.org/000/260/766/role_based_rights_in_arti_cial_social_systems.pdf [Accessed 5 January 2015].

*Bix, Brian* (1999). Jurisprudence: Theory and context. 2nd Edition. Sweet & Maxwell, London.

*Chun, Wesley* (2007). Core python programming. 2nd Edition. Prentice Hall Professional.

*Curtotti, Michael & McCreath, Eric* (2012). Enhancing the visualization of law. Paper presented at the Law via the Internet Twentieth Anniversary Conference, Cornell University. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160614 [Accessed 5 January 2015].

*Eppler, Martin J.* (2004). Knowledge communication problems between experts and managers. An analysis of knowledge transfer in decision processes. Paper # 1/2004. University of Lugano, Faculty of Communication Sciences, Institute for Corporate Communication. https://doc.rero.ch/record/5197/files/1_wpca0401.pdf [Accessed 5 January 2015].

*Eppler, Martin J. & Burkhard, Remo A.* (2004). Knowledge visualization. Towards a new discipline and its fields of application. ICA Working Paper #2/2004, University of Lugano. http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.134.6040&rep=rep1&type=pdf [Accessed 5 January 2015].

*Fox, Christopher* (2006). Introduction to Software Engineering Design: Processes, Principles and Patterns with UML2. Addison-Wesley Longman Publishing Co., Inc.

*Haapio, Helena* (2013). Next Generation Contracts: A Paradigm Shift. Doctoral Dissertation, University of Vaasa. Lexpert Ltd, Helsinki.

*Haapio, Helena* (2012). Making Contracts Work for Clients: towards Greater Clarity and Usability. In E. Schweighofer et al. (eds), Transformation of Legal Languages. Proceedings of the 15th International Legal Informatics Symposium IRIS 2012. Band 288. Österreichische Computer Gesellschaft, Wien, pp. 389–396.

*Haapio, Helena* (2006). Business Success and Problem Prevention through Proactive Contracting. In P. Wahlgren (ed.), Scandinavian Studies in Law, Vol 49: A Proactive Approach. Stockholm Institute for Scandinavian Law, Stockholm, pp. 150–194. http://www.scandinavianlaw.se/pdf/49-9.pdf [Accessed 5 January 2015].

*Harris, James William* (1980). Legal Philosophies. Butterworths & Co Publishers Ltd, London.

*Horstmann, Cay S.* (2006). Big Java. 2nd Edition. John Wiley & Sons.

*Kapitze, Cushla* (2009). Rethinking Copyrights for the Library Through Creative Commons. In Library Trends, Volume 58, Number 1, Summer 2009, pp. 95–108. The Johns Hopkins University Press.

*King, Martin Luther (Jr)* (1965). The American Dream. Speech delivered at Ebenezer Baptist Church, Atlanta, Georgia, on 4 July 1965. http://mlk-kpp01.stanford.edu/index.php/encyclopedia/documentsentry/doc_the_american_dream/ [Accessed 5 January 2015].

*Mitchell, Catherine* (2013). Contract Law and Contract Practice. Bridging the Gap between Legal Reasoning and Commercial Expectation. Hart Publishing.

*Norman, Donald* (2013). The Design of Everyday Things: Revised and Expanded Edition. Basic Books.

*Passera, Stefania, Haapio, Helena & Curtotti, Michael* (2014). Making the Meaning of Contracts Visible – Automating Contract Visualization. In E. Schweighofer et al. (eds), Transparency. Proceedings of the 17th International Legal Informatics Symposium IRIS 2014. Band 302. Österreichische Computer Gesellschaft OCG, Wien 2014, pp. 443–450.

*Patterson, Dennis* (ed.) (1996). A companion to philosophy of law and legal theory. Blackwell.

*Pohjonen, Soile* (2009). Law and Business – Successful Business Contracting, Corporate Social Responsibility and Legal Thinking. JFT – Tidskrift utgiven av Juridiska Föreningen i Finland 3–4/2009, pp. 470–484.

*Pohjonen, Soile & Visuri, Kerttuli* (2008). Proactive Approach in Project Management and Contracting. In H. Haapio (ed.), A Proactive Approach to Contracting and Law. International Association for Contract and Commercial Management & Turku University of Applied Sciences, Turku, pp. 75–95.

*Reingold, Beth* (1996). Conflict and cooperation: Legislative strategies and concepts of power among female and male state legislators. The Journal of Politics, Vol. 58, No. 2, pp. 464–485.

*Siedel, George J. & Haapio, Helena* (2010). Using proactive law for competitive advantage. American Business Law Journal, Vol. 47, No. 4, pp. 641–686. doi:10.1111/j.1744-1714.2010.01106.x.

*Stefanou, Constantin & Xanthaki, Helen* (eds) (2013). Drafting legislation: a modern approach. Ashgate Publishing Ltd.

*Thompson, Simon* (1999). Haskell: the craft of functional programming. 2nd Edition. Addison-Wesley.

# A Right to Access Implies A Right to Know: An Open Online Platform for Research on the Readability of Law

Michael Curtotti\*
Eric McCreath°

\* *Legal Counsel, ANU Students Association & ANU Postgraduate and Research Students Association, PhD Student, Research School of Computer Science, Australian National University*
° *Lecturer, Research School of Computer Science, Australian National University*

**Abstract.** The widespread availability of legal materials online has opened the law to a new and greatly expanded readership. These new readers need the law to be readable by them when they encounter it. However, the available empirical research supports a conclusion that legislation is difficult to read if not incomprehensible to most citizens. We review approaches that have been used to measure the readability of text including readability metrics, cloze testing and application of machine learning. We report the creation and testing of an open online platform for readability research. This platform is made available to researchers interested in undertaking research on the readability of legal materials. To demonstrate the capabilities of the platform, we report its initial application to a corpus of legislation. Linguistic characteristics are extracted using the platform and then used as input features for machine learning using the Weka package. Wide differences are found between sentences in a corpus of legislation and those in a corpus of graded reading material or in the Brown corpus (a balanced corpus of English written genres). Readability metrics are found to be of little value in classifying sentences by grade reading level (noting that such metrics were not designed to be used with isolated sentences).

**Keywords:** readability, legislation, legal informatics, corpus linguistics, machine learning, natural language processing, readability metrics, cloze testing

## 1. Background and Motivation

We are embedded in a network of legal rules. We are not always able to understand those rules. Sometimes social heuristics or specific training

1

(as, for example, in road rules) enable us to understand and comply with law. Often considerable expense is invested in 'explaining' the law to citizens: such as through official government information supplementing legislation, or through investment of private resources in legal services. As citizens we often need to know, and are entitled to know, the law which affects us. In a democratic context, legal rules are theoretically the outcome of consultative processes in which the entire community has a voice and in which the interests and views of the members that make it up are given due recognition and protection.

The internet has transformed the way in which society engages with legislation. It has changed how legal professionals access the law. As significantly, it has expanded and changed the audience which accesses and reads legislation. The Declaration on Free Access to Law states that public legal information is digital common property and the common heritage of mankind and calls for law to be accessible to all on a non-profit basis and free of charge.[1] This Declaration is made in the context of the considerable effort by LII's and others to achieve the practical realisation of such free access.(Martin; J., 2005)

In the UK, the Office of Parliamentary Counsel is pursuing a 'Good Law' initiative, a key objective of which is to make law more usable. The UK First Parliamentary Counsel observed:

> *Legislation affects us all. And increasingly, legislation is being searched for, read and used by a broad range of people. It is no longer confined to professional libraries; websites like legislation.gov.uk have made it accessible to everyone. So the digital age has made it easier for people to find the law of the land; but once they have found it, they may be baffled. The law is regarded by its users as intricate and intimidating.*(OPC-UK, 2013)

They note that while in the past readers of UK legislation tended to be legally qualified, that is no longer true. They report an audience of two million unique visitors per month for the legislation.gov.uk site.(OPC-UK, 2013) Similarly in the NZ case the users of legislation has broadened: *It*

---

[1] http://www.worldlii.org/worldlii/declaration/.

*seems once to have been supposed that law was the preserve of lawyers and judges, and that legislation was drafted with them as the primary audience. It is now much better understood that acts of Parliament (and regulations too) are consulted and used by a large number of people who are not lawyers and have no legal training.* There the government legislation website received 30,000 unique visitors per month.(NZ, 2008, p 14)

In 2008, the New Zealand Law Commission and the New Zealand Parliamentary Counsel's Office together undertook an inquiry into the Presentation of Law starting from the proposition that: *'It is a fundamental precept of any legal system that the law must be accessible to the public.'* Their inquiry identified three aspects of access to law: availability to the public (such as hard copy or electronic access), 'navigability' - the ability to know of and reach the relevant legal principle, and finally accessibility in the sense of the law *'once found, being understandable to the user.'* (NZ, 2008) The issues paper which preceded their report put it more succinctly:

> *Citizens should be able to know and understand the law that affects them. It is unfair to require them to obey it otherwise. This is an aspect of the rule of law.*(NZ, 2007)[2]

Concepts of 'understandability', or this third category of accessibility, are closely related to the concept of readability which is the subject of this paper. DuBay reviews a number of the definitions that are offered for readability: 'readability is what makes some texts easier to understand than others'; 'the ease of understanding or comprehension due to the style of

---

[2] Interestingly is difficult to find this principle clearly enunciated in primary sources (for example in human rights documents). An example that approaches it may be found in article 14.3 of the International Covenant of Civil and Political Rights which provides the right to be informed of charges in a language the individual understands, and the right to a free interpreter). The New Zealand Commission and Parliamentary Counsel note that in their case there is no principle of statute law that 'it must be understandable'. (NZ, 2008) Nonetheless 'understandability' is a guideline is to Departmental officers and drafters involved in the creation of legislation: *"For legislation to command public acceptance it must meet certain standards. It must be developed in accordance with proper processes, reflect legal principle, be technically effective, and be able to be understood by those to whom it applies. NZ Legislative Advisory Council Guidelines on Process and Content of Legislation"*.

writing'; 'ease of reading words and sentences' as an element of clarity; 'the degree to which a given class of people find certain reading matter compelling and comprehensible'; and 'The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting'.(DuBay, 2004) There is some variance in these definitions but they have in common (explicitly or implicitly) orientation to the needs and characteristics of a given group of readers and they assume that it is possible for a writer, by changing the selection and organisation of words, to communicate essentially the same concepts while facilitating understanding.

Kohl carries out a study of the principles of accessibility in the context of online publication of foreign laws. She notes the existence of two rationales for accessibility (including in the sense of an ability to 'know' the law). Firstly, it is unfair for a citizen to be subject to liabilities if they are unable to know the law. This rationale focuses on human and societal values. Secondly, the purpose of the law maker is to achieve compliance with law, and thus the law maker wishes it to be known. From this viewpoint, the regulator's interest in administrative effectiveness and efficiency is a motivation for ensuring access and knowledge. She notes that although legal jurists and courts propound the principle that laws should be clear or understandable as an element of the rule of law, a failure of clarity does not necessarily result in relief from legal detriment: it may amount to a *moral* principle but its effect in law is uncertain. (Kohl, 2005)

Milbrandt and Reinhardt argue for the existence of a right to access the law (in the broader sense of physical or electronic access). Principles of the rule of law, freedom of information, and principles of human rights such as the right to freedom of expression and to an effective remedy imply rights to access and know the law. Like others, they explore scenarios where access is effectively denied.(Milbrandt and Reinhardt, 2012)

A stream of action to improve the readability of law is associated with the plain language movement that particularly gathered steam during the early 1990s. Proponents of plain language cite extensive empirical studies validating the benefits of plain language for the understanding of

text. This extends to the legal context, including through widespread support of plain language measures adopted by legislative drafting offices.(Kimble, 1994) As one legislative drafting office puts it in their plain language manual:

> *We also have a very important duty to do what we can to make laws easy to understand. If laws are hard to understand, they lead to administrative and legal costs, contempt of the law and criticism of our Office. Users of our laws are becoming increasingly impatient with their complexity. Further, if we put unnecessary difficulties in the way of our readers, we do them a gross discourtesy. Finally, it's hard to take pride in our work if many people can't understand it.*(OPC-Australia, 2003)

The influence of the plain language movement has seen it mandated in both legislation and executive orders: "A number of federal laws require plain language such as the Truth in Lending Act, the Civil Rights Act of 1964, and the Electronic Funds Transfer Act. In June 1998, President Clinton directed all federal agencies to issue all documents and regulations in plain language."(DuBay, 2004)

Above we have seen both principle and practice directed to making the law more accessible in the sense of its ease of comprehension. Yet, despite this an observation made three decades ago by Bennion, the author of a leading text on statute law, could just as appropriately be made today:

> *It is strange that free societies should thus arrive at a situation where their members are governed from cradle to grave by texts they cannot comprehend.*(Bennion, 1983, p 8)

Existing empirical research on the readability of legislation supports a conclusion that legislation is inaccessible to large proportions of the population - that for many citizens it is very difficult or incomprehensible. This research moreover suggests that even plain language does not significantly alter this reality. (See discussion below in Section 3.)

The various rationales for accessibility in the sense of 'understandable' text, as discussed above, coupled with the limited progress towards its effective realization, motivates the work reported in this paper. The work is concerned, particularly from a computational perspective, with

5

identifying appropriate measures and approaches for assessing the read-ability of legislation and implementing computationally based tools for carrying out readability research on legislation. In section 2 we describe both well established and newer approaches for assessing readability including traditional readability metrics, human-centred evaluation and natural language processing and machine learning. Section 3 reviews existing research on the readability of legislation. These two sections provide a baseline for further research that might be undertaken on readability of legislation.

Section 4 describes the development of an online platform for readability research, which is offered as an open service for researchers interested in carrying out readability research. The development of this platform is part of a broader body of research on the development of computational tools for reading and writing law. [3] The platform is made available to any researchers who may wish to carry out readability research on legislative materials (or indeed any other text). The plat- form provides a number of readability tools. A tool is provided for the extraction of readability metrics from text. A second tool is designed to enable "cloze testing" (a method widely agreed to be an accurate method for measuring the readability of text). The site also provides a tool for carrying out subjective user evaluation of a text. Finally, the platform provides access to natural language processing facilities which can be used for extraction of a variety of language features such as parts of speech and ngrams. [4] The tools are accessed through a straight forward interface and are accompanied by documentation to facilitate usability.

In section 5 we report the application of this platform for initial investigations on three corpora: a corpus of graded readers, the Brown Corpus and a corpus of Australian federal legislation.

Leaving aside the theoretical justifications that might be advanced to support this view, the axiomatic position taken by this paper is that all

---

[3] For details see http://cs.anu.edu.au/people/Michael.Curtotti.
[4] An ngram is simply a sequence of a given length e.g. a bigram is a sequence of two letter, two words, or two parts of speech.

individuals subject to law are entitled to know its content and therefore to have it written in a way which is reasonably accessible to them.

## 2. Approaches to Assessing Readability

In seeking to enhance the readability of legislation, a question which naturally arises is how to assess whether given text is 'readable' or 'more readable'. Within a computational context we are particularly interested in the potential for enhancing the assessment of readability through application of computational techniques. Readability metrics naturally suggest themselves as an area of investigation, given their widespread use.

While readability metrics, such as the Flesch metric are well known (for example incorporated into Microsoft Word), their reliability and relevance are disputed both within and beyond the legislative context. Apart from such metrics, a number of other possibilities exist: user evaluation (such as comprehension testing or cloze testing and more recently crowdsourcing) and application of techniques arising from recent natural language processing and machine learning studies of readability.

### 2.1. READABILITY METRICS

Reading measures such as the Flesch, Flesch-Kincaid, Gunning, Dale-Chall, Coleman-Liau and Gary-Leary are among the more than 200 formulas which have been developed to measure the readability of text. These formulas (although varying in formulation) address two underlying predictors of reading difficulty: semantic content (i.e. the vocabulary) and syntactic structure. Vocabulary frequency lists and sentence length studies both made early contributions to the developments of formulas. The Flesch formula calculates a score using average sentence length and average number of syllables per word as measures for determining text difficulty. Formulas of this kind are justified on the basis of their correlation with reading test results. For example, the Flesch formula correlated at levels of 0.7 and 0.64 in different studies carried out in 1925 and 1950 with user tested texts.(DuBay, 2004)

The uses and abuses of such formulas have been widely debated. An important observation in this context is that these tests were not conceived as measures of comprehensibility of text, rather they were designed to help teachers select appropriate texts for children of different ages.(Woods et al., 1998)

In 1993 an Australian Parliamentary Committee report on clearer legislation (having reviewed use of readability metrics) commented:

> *Testing for the readability of legislation by using a computer program is of limited value. The most effective way of testing legislation is to ask people whether they can understand it - a comprehension test. Ideally this type of testing should occur before the legislation is made.* (Melham, 1993)

Evidence presented to the Inquiry included the view that research had undermined the validity of readability metrics and the view that readability metrics could mislead by mis-categorising the complexity of legislative sentences (Melham, 1993, p. 98).

A review of methods for measuring the quality of legislation carried out in New Zealand observed that readability metrics can only play a limited screening role in the prediction of readability. It considered such metrics to have limitations such as not detecting how complex ideas are, whether the language is appropriate to the audience or whether a sentence is ambiguous. They note that legislative drafters in the UK have concluded that such tests do not measure readability in a comprehensive sense, but that they seem reasonably good as an initial indicator of problematic text.(PCO-NZ, 2011)

Despite their limitations, readability metrics are used in practice and have a body of supporting research. They have been influential and continue to be widely used:

> *Writers like Rudolf Flesch, George Klare, Edgar Dale, and Jeanne Chall brought the formulas and the research supporting them to the marketplace. The formulas were widely used in journalism, research, health care, law, insurance, and industry. The U.S. military developed its own set of formulas for technical-training materials. By the 1980s,*

8

*there were 200 formulas and over a thousand studies published on the readability formulas attesting to their strong theoretical and statistical validity* (DuBay, 2004).

A debate carried out between a readability specialist, computer scientists and others in the context of computer documentation is illuminating as to the limitations of readability metrics. Klare, the readability specialist participating in the debate, cited a number of limitations of readability metrics. These included that they function best as screening devices only, need to be interpreted in light of reader characteristics, cannot be used as formulas for writing style 'since changes in their index variables do not produce corresponding changes in reader comprehension' and should be used in conjunction with other approaches such as use of human judges, cloze procedure and usability testing. Further, readability metrics are designed for larger blocks of text providing a connected discourse and won't work well on disconnected fragments or single sentences (something relevant to the experiments reported below).(Klare, 2000) Others note the poor correlation between different readability metrics themselves.(Woods et al., 1998) Beyond this, some studies have found poor correlation between human judgements as to readability and the scores assigned by readability metrics(De Clercq et al., 2013; Harrison and McLaren, 1999; Heydari and Riazi, 2012). Heydari et al. observation perhaps sums up the state of research:

*If any conclusion is possible to draw from the hodge-podge of studies done on readability formulas, it is that there are two opposite views toward the use of them. Both of these two views have been advocated by different researchers and there is enough empirical evidence for each to be true. Thus, it can be declared openly that the formulas have both advantages and disadvantages.* (Heydari and Riazi, 2012)

With such conclusions, some caution is required in using readability metrics. The caution is reinforced in respect of legal language, particularly legislative language. Little validation has been undertaken of readability metrics in the context of legal language. Until that validation is carried out and the parameters of valid application understood, any conclusions based on application of such metrics must be qualified with uncertainty. Their advantage is that they are readily calculated without significant investment of human resources - a factor that has likely

9

contributed to their widespread use. The Readability Research Platform includes tools for extracting various readability metrics.

## 2.2. COMPREHENSION TESTING, CLOZE TESTS AND CROWDSOURCING

In this section we review some human centred approaches to evaluating the readability of text. Such methods equate to the field of user evaluation, in human computer interaction. Such methods are perhaps the most promising for application to improving the readability of legal language. If properly implemented, such tests can measure how understandable text is to readers, and can be targeted to particular reader groups of interest (e.g. the general public or individuals particularly affected by an item of legislation). Their disadvantage is that they are resource intensive to carry out, while crowdsourcing requires access to platforms with large user traffic and programming skills.

### 2.2.1. *Comprehension Testing and User Evaluation*

A traditional method of testing the ability of a reader to understand a text is to administer a comprehension test. This method can be used in reverse to assess the difficulty of the text, for given populations of readers. Tests are deployed by having a student read a passage and then answer multiple choice questions regarding its content.(DuBay, 2004)

### 2.2.2. *Cloze Tests*

The cloze procedure involves testing the ability of readers to correctly reinsert words that have been deleted from a given text. Typically the test is administered by deleting every *n*th word in the text. When used to assess the readability of a text the cloze procedure is administered by deleting every fifth word (including sometimes five different versions of the text staggering the deletion), and replacing it with a blank space, which the reader must fill in by guessing the missing term (Bormuth, 1967). Although initially conceived as a remedy for the shortcomings of readability formulas, the cloze procedure came to complement conventional reading tests (DuBay, 2004). Cloze procedure was also developed to provide a more valid measure of comprehension than traditional multiple choice comprehension tests.(Wagner, 1986) Of greatest interest in this context is use of cloze tests as a measure of the readability of a text. Bormuth notes that there is a high correlation

between cloze readability testing and comprehension testing on human subjects:

> *A reasonably substantial amount of research has accumulated showing that cloze readability test difficulties correspond closely to the difficulties of passages measured by other methods.* (Bormuth, 1967)

Bormuth cites studies, including his own, which show correlations ranged from .91 to .96 with the difficulty of texts assessed with traditional comprehension tests.(Bormuth, 1967) When properly applied the cloze test provides an indicator of how difficult a text was for given readers. A cloze score of below 35% indicates reader frustration, between 35% and 49% is 'instructional' (the reader requires assistance to comprehend the material) and 50% or above indicates independent reader comprehension.(Wagner, 1986)

As we see below (section 3), the cloze procedure has been used as a means of assessing the readability of legislation. The Readability Research Platform described below includes a cloze tool, which is in demonstration phase.

### 2.2.3. *Crowdsourcing*

The emergence of large populations of online users, opens the possibility of such users being engaged in the task of assessing the readability of legislation. A parallel might be drawn with crowdsourcing used to support scientific research such as through the Zooniverse platform, some projects of which use human judgements to support the classification of images of galaxies, to cite one example.[5] De Clercq et al. undertake an evaluation of the effectiveness of crowdsourcing as a method of assessing readability. They compared the accuracy of crowdsourced human judgements of the readability of text with those of expert judges, finding a high level of agreement in readability ranking between the experts and crowdsourced users. crowdsourced users were presented with two randomly selected texts of one to two hundred words and invited to rank them by readability. Expert teachers, writers and linguists were given a more complex task of assigning a readability score to each presented text. In addition to concluding that crowdsourced user judgements and expert judgements were highly correlated as to readability ranking, they found

---

[5] How Do Galaxies Form Classification Project https://www.zooniverse.org/project/hubble.

that readability metrics had a lower correlation with these two judgement sets.(De Clercq et al., 2013)

A more general study by Munro et al. on the use of crowdsourcing in linguistic studies concluded that there was a high correlation between traditional laboratory experiments and crowdsourced based studies of the same linguistic phenomena. Among their conclusions was that crowdsourced judgements closely correlated with cloze testing results, which as we have seen above is a key approach to undertaking readability studies. (Munro et al., 2010) We are unaware of any studies which have used crowdsourcing to assess the readability of legislative text. There does not seem to be any serious impediment to using such an approach and the Readability Research Platform includes a demonstration tool for collecting user evaluations of text.

## 2.3. MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

Recent years have seen a growing body of research seeking to apply natural language processing and machine learning to assessing the readability of text. The term 'natural language processing' represents the capacity of computers to hold and analyse large bodies of text. Natural language processing can be applied to represent text as collections of characters, collections of words, to annotate words with their grammatical type (such as noun, verb, adjective etc.), to aggregate words into grammatical phrases and to represent the syntax of sentence as a grammatical tree. Such purely functional annotation can be extended to information extraction - the identification of entities such as persons, organisations, places etc, and the identification of relationships. Such work falls under the heading of natural language processing.

Machine learning is grounded in mathematical theory and provides well elaborated processes of enabling patterns to be learnt from a given body of data. Data (for example linguistic data) is represented as a set of 'feature', 'value' pairs associated with each item from the dataset. For example a sentence has associated with it a set of features such
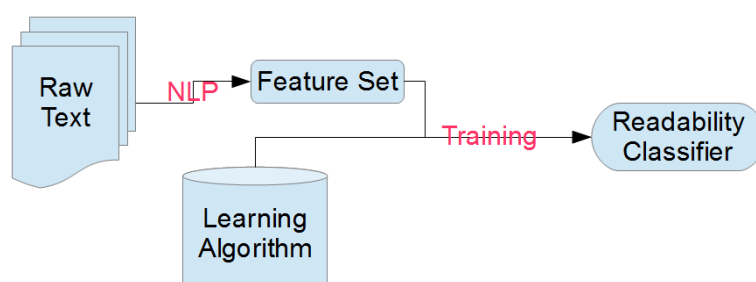
Fig. 1. A typical natural language processing and machine learning pipeline in application to readability

as its length, the number of words, the parts of speech of those words, the given vocabulary and patterns such as the occurrence of two words in sequence. Such features can then be used to learn a model which with a known level of accuracy predicts (for example) the classification of a previously unseen sentence. Machine learning includes both 'supervised' and 'unsupervised' learning. In supervised learning a data set already labelled with the appropriate classifications is provided as input to the learning algorithm. In the unsupervised case the machine learning is carried out on unlabelled data.[6]

Readability research has applied both these processes to seek to automatically predict the readability of given text. A pipeline of trans-formations are carried out on a dataset consisting of input documents (which need be no longer than a single sentence) with the aim of learning a capacity to predict the readability of given text. Figure 1 illustrates a typical process, the desired end result of which would be a learned classification model with the capacity to correctly classify text for its readability with a known level of accuracy.

Many have in common the hypothesis that 'deeper' language features provide valuable data for the task of assessing the readability of text.

---

[6] See Bird et al. for a very accessible and practical introduction to natural language processing. Chapter six also introduces machine learning in application to the classification of text.

An exhaustive review of the application of these techniques to readability is not carried out here but a number of aspects of particular interest are highlighted. A key question is what features might assist us in assessing readability? Studies have systematically examined sets of features for their utility in assessing readability. The most straight forward features examined have been readability metrics themselves and 'surface' features such as average sentence length, average word length and average syllable length, capitalisation, punctuation. Other features studied include lexical features such as vocabulary and type/token ratio, [7] parts of speech frequencies, ratio of content words to function words, distribution of verbs according to mood, syntactic features such as parse tree depths, frequency of subordinate clauses, ngram language models, discourse features, named entity occurrences, semantic relationships between entities and anaphora occurrences. (Dell'Orletta et al., 2011; Kate et al., 2010; Feng et al., 2010; Si and Callan, 2001)

Collins Thompson and Callan in 2004 undertook a study of the use of 'language models' to predict reading grade. They build a model of grade language based on the probability of a word for each grade level. This approach was based on the observation that the probability of a word occurring in a text varies depending on the grade level of the text. However the authors were guarded in the conclusions they felt able to draw as to the effectiveness of their approach (Collins-Thompson and Callan, 2004).

Schwarm and Ostendorf in 2005, also used a language modelling approach, in combination with other features. They apply a support vector machine algorithm to undertake machine learning using features such as readability metrics, surface features, closeness of match for language models built on graded reading material, parse tree heights and number of subordinating conjunction. Their support vector machine grade prediction outperformed the Flesch-Kincaid grade measure and the Lexile measure by a wide margin. None of the features they used stood

---

[7] A 'type' is say the word 'red' and a token is any word. So in the phrase "the cat sat on the mat" the type to token ratio is 5/6, as the word 'the' occurs twice.

out as critical to classification, but removal of any degraded performance.(Schwarm and Ostendorf, 2005)

Heilman et al. in 2008 test a number of machine learning algorithms using unigram language models and full and sub-tree features as grammatical input. They attain an accuracy of 82% in predicting grade level of documents in their corpus using a combination of language features.(Heilman et al., 2008)

Pitler and Nenkova also in 2008 use adult reading materials from the Wall Street Journal graded as to readability by human judges. They note that 'readability' assessments are dependent on audience and note that graded readers designed for language learners are not generalisable to the question of general readability of more standard texts. They assess various features for predicting readability using this labelled corpus. Surface, syntactic, lexical cohesion, entity grids and discourse relations. They identify discourse relations as most predictive of readability (correlation of .48), followed by average number of verb phrases, followed by article length. Combining the various features they examined attained the highest accuracy of around 88%. Surface features (which underlie most readability metrics) they find to be poor predictors of readability.(Pitler and Nenkova, 2008)

Feng et al. undertake a study of similar scope to Schwarm noted above. Again using a corpus of graded material they seek to identify factors most predictive of readability. They find parts of speech features (particularly nouns) to be highly correlated with grade level. They also note that among surface features used in traditional readability metrics, average sentence length has the highest predictive power.(Feng et al., 2010)

Kate et al., like the Pitler study, use a labelled dataset of adult reading materials. The dataset of 540 documents is labelled by expert and naive human judges. The machine learning algorithm is then trained to predict readability from a training set labelled with expert judgements. The authors find that using diverse linguistic features, they are able to exceed the accuracy of naive human judges as to readability. As with other studies combining features produced the highest levels of accuracy.(Kate et al., 2010)

15

Aluisio et al. also apply machine learning and like other studies find that combining linguistic features increases accuracy of prediction. They are also concerned to leverage readability assessments for the task of simplifying text. (Aluisio et al., 2010)

Of particular interest for classifying the readability of legal rules are readability studies which focus on classification of single sentences or shorter text fragments. As legal rules are often written as single sentences may be of greater assistance than readability measures which focus on paragraphs or blocks of text. Dell'Orletta et al. carry out readability assessment at both document and sentence level, undertaking a binary 'hard' vs. 'easy' classification of Italian texts. As with other studies they examine a wide range of features. However they also are particularly interested in assessing features that might later be applied to the process of text simplification. Base features (such as underlie readability metrics) show little discriminative power for sentences, but they find that the addition of morpho-syntactic and syntactic features increases accuracy of sentence level classification to 78%.(Dell'Orletta et al., 2011; Sjoholm, 2012)

Sjoholm's 2012 thesis also addresses predicting readability at sentence level. He notes the absence of existing metrics for predicting readability at sentence level. He builds on previous studies by developing a probabilistic soft classification approach that rather than classifying a sentence as 'hard' or 'easy' gives a probability measure of membership of either class.(Sjoholm, 2012)

The application of natural language processing and machine learning to the task of predicting readability has made considerable progress over the last decade or so. Studies such as those above have demonstrated that prediction of readability can be significantly improved by incorporating higher level linguistic features into predictive models. Further, of interest to us, the Dell'Orletta and Sjoholm studies underline the inadequacy of traditional readability metrics (as they are based on surface features) for assessing readability at sentence level. It is also notable that only initial steps have been taken to apply findings in this field to identifying reliable methods of improving readability.

Natural language processing and machine learning, as suggested by the progress of recent research, offers considerable promise that it may allow progress in understanding and addressing readability issues in legislation. Significant is still required to adapt the existing research to application to readability in the legislative field. A limitation of such methods is that without a considerable body of labelled data, it is difficult to attain high levels of accuracy with machine learning. Obtaining reliably labelled data is best achieved through user studies of the kind described in Section 2.2. Another challenge inherent in machine learning is determining those 'features' which are most associated with readability. The work reported above provides some guidance as to which features may prove useful.

### 3. Empirical Research on the Readability of Legislation

In section 1 we noted the extensive attention given to readability of legislation by government agencies and the plain language movement. Readability is a standing concern of legislative drafting offices with plain language being a frequent goal or commitment of such offices. (Kimble, 1994; OPC-Australia, 2003) Here we seek to summarise the findings of empirical research which directly assesses the readability of legislation. Such empirical studies are limited in number and scope, though considerable work has been undertaken on tax legislation.

An early example was a study reported in 1984 in which cloze testing was undertaken on several samples of legal text including legislative language. 100 generally highly educated non-lawyers (28% had undertaken some postgraduate training) were tested. The group averaged 39% accuracy, a result close to 'frustational' level for cloze testing. Ten participants who had only high school education experienced greater difficulty, averaging 15% − a result consistent with total incomprehension.(Benson, 1984)

In 1999, Harrison and McLaren studied the readability of consumer legislation in New Zealand, undertaking user evaluations, including the application of cloze tests. They seek to answer a number of questions including: how comprehensible to consumers and retail workers is New Zealand's consumer legislation? The study found traditional readability metrics to be unreliable. The results of cloze testing on extracts from the legislation led to the conclusion that the legislation would require

17

explanation before being comprehended at adult level. For young adults (aged 18-34), comprehension levels were even lower (within the frustrational level). Paraphrase testing, where participants were asked to paraphrase the legislation, also showed that participants found the Act difficult to understand with one section proving almost impossible to access. Participants complained of the length of sentences and most felt there was a need for some legal knowledge to understand the text. All felt the text should be made easier. The researchers also inferred from cloze testing that simpler terms were required in the legislation to make it more accessible to the public.(Harrison and McLaren, 1999)

In the early 1990's Australia, New Zealand and the United Kingdom pursued tax law simplification initiatives which involved rewriting at least substantial portions of tax legislation. The goal in Australia's case was stated to be to 'improve the understanding of the law, its expression and readability'. Cloze testing on a subset of the work was however inconclusive, finding participants found both the original language and the rewritten language difficult.(James and Wallschutzky, 1997) Smith et al., reviewing the effectiveness of the same program, concluded that results fell 'far short of an acceptable bench-mark'. They used the Flesch Readability Score as a measure of readability finding that readability of sections of tax law replaced in the tax law improvement program, improved on average from 38.44 to 46.42 - a modest improvement. The result is well short of the general Flesch benchmark of 60-70 for readability. i.e. even after improvement, the legislation remained difficult to read. Over 60% of the revised legislation remained inaccessible to Australians without a university education.(Smith and Richardson, 1999) A similar study of the readability of goods and services tax legislation in Australia also applying the Flesch Readability Index, finds an average readability of 40.3 (i.e. low). Again such results exclude considerable proportions of the Australian community.(Richardson and Smith, 2002)

A study in Canada carried out usability testing on plain language and original versions of the Employment Insurance Act. Members of the general public and expert users were recruited to carry out testing. All participants completed more questions in the plain language version. Similarly all participants using the plain language versions were more accurate in their answers. All respondents, particularly those from the general public, found navigation and comprehension difficult irrespective

of version. They also found that for all versions respondents faced difficulty in understanding the material. These findings indicated that in this instance while plain language reduced difficulty it did not eliminate it. Nonetheless participants preferred the plain language version and found it easier to use.(GLPi and Smolenka, 2000)

Tanner carried out empirical examination of samples of Victorian legislation, assessing them in light of plain language recommendations of the Victorian Law Reform Commission made 17 years earlier. The authors noted that the Law Reform Commission had recommended that *on average* sentences should be no longer than 25 words and that complex sentence structure was to be avoided. In a study of six statutes they found that the average sentence length was almost double that recommended by the Commission, and that over time sentence length had increased. In the Fair Trading Act (a piece of legislation of general importance to citizens), they found that the number of sentences with six or more clauses was particularly high. Although they also note improvement in some areas, they conclude: "The net result is that many of the provisions are likely to be inaccessible to those who should be able to understand them. This is because the provisions 'twist on, phrase within clause within clause'."(Tanner, 2002)

An empirical study of the usability of employment legislation in South Africa also found that respondent accuracy improved considerably with a plain language version of the legislation. The respondents who were drawn from year 11 school students averaged a score of 65.6% when tested on the plain language version, whereas the control group scored an average of 37.7%. Like other studies it found that plain language improved comprehension.(Abrahams, 2003)

A 2003 review of the Capital Allowances Act in the UK which was rewritten as part of the UK's tax law improvement program undertook interviews with a number of professional users. These professionals in general responded that the new legislation was easier to use and more understandable.(OLR, 2003)

A similar review of the Income Tax (Earnings and Pensions) Act also carried out in the UK again found that the interviewed group (primarily tax professionals), were largely positive about the benefits of the

simplification rewrite, expressing the view that the revised legislation was easier to use and understand, although also noting the additional costs of relearning the legislation.(Pettigrew et al., 2006)

A 2010 study of the effects of the tax law simplification in New Zealand employed cloze testing to determine the degree to which the simplification attained its goals. They cite a 2007 Australian study by Woellner et al. which using cloze procedure, found that novice users of both original and amended versions did not achieve benchmark comprehension but found the new legislation (ITAA 1997) marginally easier (35% vs 24%). In their own study they reported that most of their respondents (mainly respondents unfamiliar with the tax system) found the cloze testing either difficult or extremely difficult. They found that the older (unamended) Act was the least difficult - a finding contrary to their expectation given prior research in New Zealand - this they attributed to the nature of the selections from the older legislation. The overall average cloze results was 34.17, with unfamiliar respondents achieving 30.86%. They note that less than 25% of their subjects were able to exceed the instructional guideline of 44%. (Sawyer, 2010)

The empirical readability research points to two conclusions. Firstly writing in plain language assists comprehension of legislation. Secondly legislation is generally incomprehensible or difficult to read to large sections of the population, even in those cases where plain language revision has been undertaken.

## 4. An Open Online Platform for Readability Research

### 4.1. Motivation and Description of the Platform

The previous sections of this paper provides an overview of the body of knowledge which provides context for the Readability Research Platform, which is maintained on an Australian National University server accessible via the internet[8] and which is described below. Its particular purpose is to enable an extension of the reported research on readability of legislation (and other texts for that matter), initially to meet the needs of

---

[8] http://buttle.anu.edu.au/readability/.

the authors, but later as an effort to make relevant tools available to other researchers. In this context, a number of factors contribute to the design of the tool:

- The primary use case for which the platform is designed is carrying out readability research (including on legislation).

- Given this, the platform needs to facilitate or enable the application of various readability approaches. It thus includes tools that cover the various approaches discussed above. It is also extensible, as additional tools can readily be added as need arises. The availability of these tools in one place facilitates comparative studies of different approaches, as well, it is hoped, as facilitating comparison of work undertaken by different researchers using the tool.

- The community interested in the readability of law is a multidisciplinary one. In this context the platform would preferably be accessible to researchers with little or no experience of programming. For this reason the protocols adopted in the platform are as simple as possible, avoiding frameworks that require familiarity with particular representations of data. The tool accepts plain text as its primary form of input and seeks to simplify the steps required to extract data.

- Given the scale of legislative data, the platform be capable of handling either large documents or a large number of smaller documents at a practical speed.

- The platform would ideally enable researchers to build on existing research, making it important to incorporate access to natural language processing tools, which are at the cutting edge of readability research.

- The design of the tool should enable collaboration with interested researchers through potential for integration with online legislative sites.

- The tool would ideally facilitate the reproduction of existing results in the readability field.

Apart from its use for research, the demonstration pages on the website provide visual introductions to the readability tools they demonstrate.

Where available, the platform makes use of existing open access libraries for carrying out underlying natural language processing, while abstracting away details of use of these packages in application to readability tasks. Natural language processing is provided by either the NLTK Language Toolkit or Montylingua.(Bird et al., 2009; Liu, 2004) Most readability metrics are extracted using a plug in to NLTK developed by Thomas Jakobsen and Thomas Skardal. http://code.google.com/p/nltk/source/browse/trunk/nltk_contrib/nltk_contrib/readability/



Fig. 2. The Readability Research Platform Website

## 4.2. USING THE READABILITY RESEARCH TOOL

The site provides a number of demonstration pages illustrating the kinds of outputs that can be extracted using the platform (see Figure 2). These include: readability metrics, natural language processing, cloze testing and user evaluation. A help page is provided which is designed to address the needs of researchers. The page describe commands that can be sent to the server which returns either data extracted from text provided as input or html (that can be used as a widget in another web page). These tools are intended primarily for the purpose of data extraction from text. Data

that can be obtained includes readability metrics, surface features, parts of speech, chunk phrases and ngram data. The data is returned as text which can either be saved to file or used as input to code developed by the researcher.

The server will respond to a http request sent to the server in for- mats described on the help page. Also the server functionality can be explored manually using the browser's url address box. For example typing: http://buttle.anu.edu.au/readability/?getariXXXXThe brown fox is quick.', and sending it to the server, will return the ARI readability metric for the sentence: 'The brown fox is quick.' A list of available commands and their descriptions is provided at the website help page.

The primary scenario for which to the platform is designed is automated extraction of data from text. While it is possible for a researcher to cut and paste text into the tool, this is impractical in most real world research scenarios. In order to retrieve data the researcher can use simple scripts which send http requests to the server and retrieve the requested data. The retrieval of data can be achieved in a few lines of code. The key steps in a typical use case scenario are:

1. create a local file into which to save results;

2. send a command (any arguments) and the text to be analysed to the server;

3. save the response from the server to the local file;

4. analyze resulting data using an external statistical package.

Two examples of simple scripts written in Python are provided in Appendix A which illustrates these steps. If the resulting data is comma delimited and saved into a file with a .csv extension, it can be opened in Microsoft excel and analysed or subjected to further processing.

A more complex example of use of the Readability Research Platform is provided in Appendix B. The consists of the calls made in the iPython command line interface, a script and a class for saving data into the Weka Machine Learning Software data format 'ARFF'. The example in Appendix B, which is written in Python, can be replaced with code

written in another programming language. The resulting datafile could then be used for carrying out machine learning using Weka package.

## 4.3. TESTING AND PROFILING

Unit testing was carried out on individual metrics to ensure the code behaves as intended. The Selenium testing platform was used for these tests, which confirmed the accuracy of a number of readability metric results on short input texts.

Also performance profiling was completed on a variety of the natural language related commands to understand and compare their performance characteristics. This was done by providing the server with a document and timing how long the server took to complete the test for a variety of different configurations. The documents had word counts ranging from 100 to 1000 in increments of 100. The results are graphed and shown in Figures 3 and 4.

The graph in Figure 3, using a logarithmic scale, shows the large range in performance for different processing tasks. Extraction of British National Corpus Metrics (which was slowest) took in the order of 10s of seconds, whereas the simple ARI metric takes tenths of a second to process on similar sized documents.
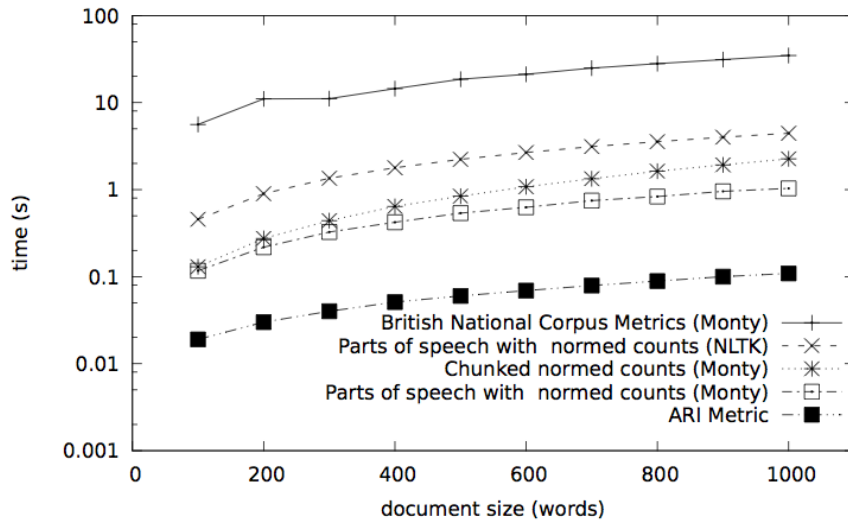
24

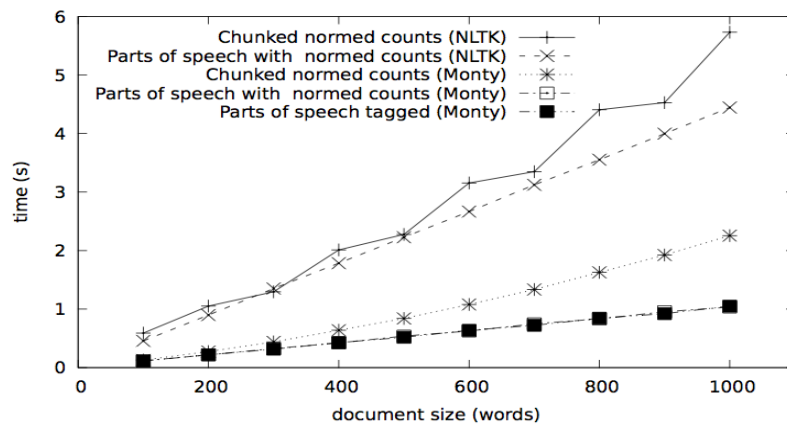Fig. 3. Log Time Performance of Selected Data Extraction Commands by Document Size



Fig. 4. Scaling of Performance by Document Size

25

The graph in Figure 4 shows that the parts of speech processing are linear with respect to performance. This would suggest these evaluations would be viable for large documents. Note that the Montylingua tool performed better than NLTK for the processing parts of speech by a factor of approximately 4.3. Also from this graph it is clear that the chunking code contains some quadratic scaling, this indicates the evaluation may be problematic if the documents become very large. There was little difference in performance between raw or normed counts so we have only graphed the normed count versions.

The speed of the platform, although far from instantaneous, is sufficient for a wide range of realistic research scenarios. For example extracting parts of speech counts for a 1,000,000 word corpus using the NLTK option (one of the slower commands) would take about an hour and a quarter. A significant factor in performance is the inherent computational complexity of tasks such as parts of speech tagging which are likely to already be optimized in the underlying code. Nonetheless, we have undertaken little work to optimize performance, a task that could be pursued as the platform is further developed.

## 5. Initial Investigations of Legislation and Readability using Machine Learning

The Readability Research Platform described above was used, through its http request protocols, to undertake initial investigations to characterise legislation for readability purposes. The focus of investigation was at the level of individual sentence or individual legal rule (the latter often constituting a single sentence in drafting practice). This enables us to investigate legislative language from the point of view of the citizen or user seeking to understand an individual rule or sentence.

We investigated a number of questions.

1. Do traditional readability metrics or surface features of a sentence assist us in assessing the readability of the sentence?

2. Does parts of speech or chunk data from a sentence assist in assessing its readability?

26

3. Do features such as the above provide us with a measure of whether legislative 'sentences' are 'normal' English?

Three corpora of English language were used to investigate these questions.

- A corpus of extracts from graded readers which was downloaded from the internet (graded reader corpus).[9]

- The Brown University Standard Corpus of Present-day American English which is a balanced corpus of English genres.(Francis and Kucera, 1964) The corpus is available through the Natural Language Toolkit.(Bird et al., 2009)

- A corpus of 'popular' legislation, identified as such on the official Australian legislation website (www.comlaw.gov.au), which was downloaded from that site and from the AustLII website (austlii.edu.au) and compiled into a corpus of legislation. Head material and appendices and notes were removed from the legislative corpus as such material does not form part of the legal rules themselves.[10]

## 5.1. DO READABILITY METRICS AND SURFACE FEATURES ASSIST IN ASSESSING THE READABILITY OF A SENTENCE?

The Readability Research Platform[11] was used to extract readability metrics and "surface features" from individual sentences from the graded reader corpus. The resulting data file was in 'ARFF' format, and was used to carry out machine learning using the Weka Data Mining Software Package.(Hall et al., 2009) 'Classification' was used to explore how useful

---

[9] http://www.lextutor.ca/graded/. A copy of the graded corpus used in this research can be obtained at http://cs.anu.edu.au/people/Michael.Curtotti/data/gradedcorpus.zip.
[10] http://cs.anu.edu.au/people/Michael.Curtotti/data/legislativecorpus.zip.
[11] http://buttle.anu.edu.au/readability/.

the extracted features (in this case readability metrics and surface features) were for classifying the material into their correct grades.

Readability metrics are typically designed for use on passages of text of 100 words or more (as we discussed above). Even though they are not designed for the task of assessing readability of individual sentences, are they nonetheless useful?

The potentially limited value of such metrics for readability assessments at sentence level is illustrated by Figure 5, which was generated by the Weka machine learning package on data extracted from the Graded Reader Corpus. Each colour represents a distinct grade level, showing the distribution of Coleman Liau Index results for sentences for that grade. The extensive overlap of the metric's results for the different grades will be evident. The implication is that if all that is known about a sentence is its Coleman Liau Index, it will be very difficult to say which grade it comes from. Although the mean for the Coleman Liau distribution can be seen to move higher as the grade level increases, each grade level has a very similar range. This overlapping distribution is typical of what we observed with respect other readability metrics.
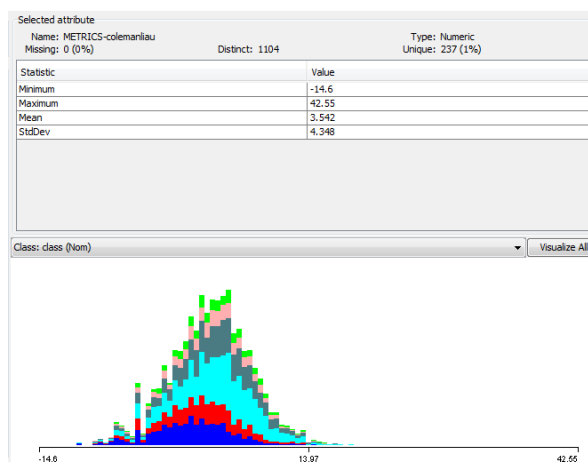


Fig. 5. Stacked Histogram Distribution Visualization of Coleman Liau Metric for Six Grade Levels from Graded Reading Corpus

28

We carried out multiclass classification on 14456 data items trialling a number of learning algorithms. The baseline accuracy value of 22.2% (ZeroR − i.e. guessing the most frequent class) was increased to 28.4% accuracy in the case of the Weka package support vector machine implementation (SMO) tested using ten-fold cross validation. The highest accuracy was 36% on any classification for any particular grade. By themselves, readability metrics are insufficient for the task of distinguishing reading grade level, at sentence level. Such metrics are not completely useless at sentence level either, however, as accuracy over the base level was increased by 6.2%.


5.2. Does parts of speech or chunk data from a sentence assist in assessing its readability?

Language may also be analysed by parts of speech (POS) (such as determiners, nouns, verbs, prepositions), and by phrase chunks (noun phrases, verb phrases, adjectival phrases and prepositional phrases).

The language features provided by POS and chunks, is additional to that provided by readability metrics. Do such features enhance classification of sentences by grade level?

We found that machine learning using these features alone, or these features in combination with readability metrics and surface features, does enhance the classification of sentences according to grade reading level.

Tests were carried out on a smaller set of 1613 data points drawn from the graded reader corpus with additional features and then machine learning classification was carried out using ten fold cross validation.

The baseline ZeroR accuracy was 19.9%. Machine learning using just parts of speech and chunk information increased accuracy to a maximum of 30.4%, using Bayesnet learning. Using parts of speech, chunking information and readability metrics and surface features as well as ranking and frequency information from the British National Corpus, increased accuracy to a maximum of 35.2%, using the Decision Table algorithm. Again ten fold cross validation was used for machine learning. In no case was accuracy on any particular grade higher than an F-measure

of 0.44. Accuracy increased by 15.3% over the base- line. Again we see that even with the additional features, classification results remain poor.

A qualifier with this particular trial is the significantly smaller number of data points used for the machine learning.

## 5.3. Do Readability Metrics allow us to reach conclusions as to whether legislative 'sentences' are 'normal' English?

Above we saw that readability metrics and surface features provide limited capacity to determine if a sentence belongs to a particular grade level. By contrast the same is not true of the ability to distinguish sentences drawn from legislation from other English sentences.

Legislative sentences, as characterised by readability metrics and surface features, are quite distinct from the graded reader material as illustrated by a visualization of a number of these metrics. In Figure 6 for each metric, legislative sentences (the top row in tan) are an outlier. The figure show the Weka summary visualization of the distribution of values for some of these metrics and the 'words per sentence' surface feature. From visual inspection it can be seen that the distribution of these metrics for each of the graded readers is similarly distributed, whereas legislative sentences have a much broader range of values.
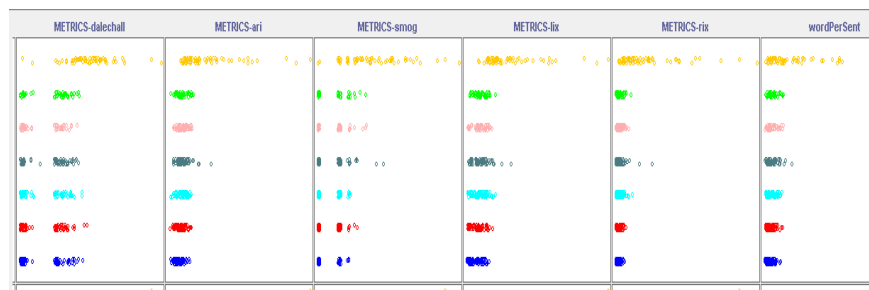


Fig. 6. Distributions of Metrics for Graded Reading Material and Legislation. The top row shows range of values for legislation for illustrated metrics, lower lines illustrate relative distribution ranges for graded readers.

The hypothesis suggested by this visualization is that legislation is significantly different from normal English usage. We may further hypothesise that this difference may contribute to reading difficulty for readers expecting to find 'normal English. Such a hypothesis would be consistent with the findings of studies that we have examined above that legislative texts are often inaccessible to non-professional readers.

The hypothesis suggested by the visualization is further supported by machine learning which we carried out on both the legislative corpus and the graded readers. Machine learning is far more effective at distinguishing legislative sentences from the graded readers. A balanced and randomized dataset was prepared which included both legislative sentences and sentences from the graded reader material. The dataset contained a total of 16 566 items. The ZeroR default accuracy was 17.9%. On this dataset machine learning algorithms increased accuracy to 30.7% (JRip), 34.4% (REPTree), 34.5% BayesNet, 34.9% (SMO), 34.1% (Decision Table) and 33.1% Naive Bayes. As with the Brown corpus comparison discussed below, the F-measure accuracy of classification of legislation was considerably higher than for readability grades: 0.87, 0.89, 0.79, 0.83, 0.83 and .80 respectively for the different learning algorithms. 0.37 was the highest F-measure accuracy for the classification of any grade level on any of the learning algorithms used.

A potential objection to the validity of this comparison is that the graded readers are not in themselves 'normal' or real world English. Especially at lower grade levels, the readers are simplified English produced for the purpose of assisting readers to develop their reading skills. A comparison is required with real world English.

To address this objection we also carried out a further comparison using the Brown Corpus which is a balanced corpus of different genres of English text: i.e. it is a representative sampling of the major forms of written English. Given that the Brown corpus is not organised by assumed difficulty of reading, we would expect that readability metrics would not be particularly useful in distinguishing different genres (not being designed for this task).

Again visualization (Figure 7) suggests that legislative sentences are an outlier. There is in this case more variance between the Brown Genres,

nonetheless legislative sentences have a much wider range of variation for readability metrics and surface features as compared to the genres.

The test carried out on the corpus confirmed this with JRip machine learning using readability metrics and surface features only increasing the base ZeroR figure from 9% to 10%. This result also allows a conclusion that the kinds of features that readability metrics provide are unable to
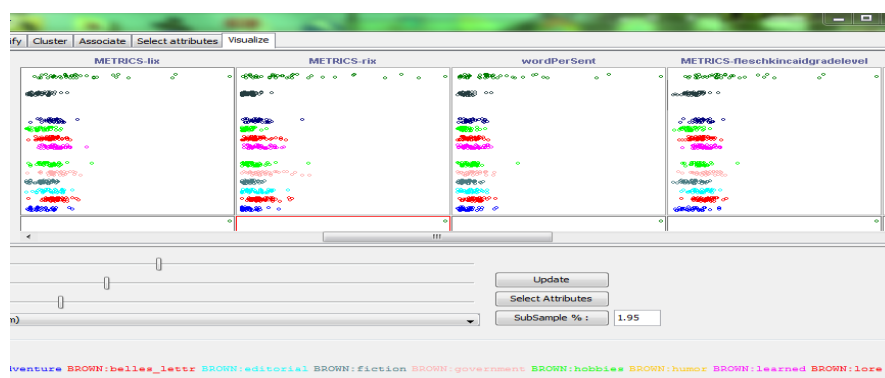


distinguish between genres of English at a sentence level.

Fig. 7. Distributions of Metrics for Brown Genre and Legislation (the top row is Legislation). As with Figure 6 lower rows show relate metric value distribution, but in this case for Brown genes.

Testing with legislative sentences versus Brown genres are not as marked as the results with graded reading material, but nonetheless legislative sentences are the most distinctive genre by a large margin if compared with the genres in the Brown corpus. Whereas the F- measure for classifying Brown corpus genres does not rise above 0.17, for legislation the figure rises to 0.47, with a precision of 73% and a recall of 35%. The comparison with a balanced corpus of written English increases confidence that legislative language is indeed 'different' as far as readability metrics and surface features are measures of that difference.

Initial work was also undertaken to examine whether other features (parts of speech and chunk data), also suggest a significant difference in legislative language. A further set of experiments was undertaken

32

analysing a smaller dataset of Brown genres and legislation consisting of 3691 datapoints. JRip in this instance produced unreliable results as it dealt with legislation as a residual category into which otherwise unclassified items were labelled.

A number of different learning algorithms were therefore applied. Apart from JRip (and Conjunctive Decision Table, which also produced low results (11% overall accuracy)) each machine learning algorithm found it considerably easier to correctly classify legislative sentences as opposed to sentences from Brown genre categories, using parts of speech and chunk phrase data. (See Table I)

| Machine Learning Algorithm | F-Measure Accuracy Legislative Sentences | Nearest or Highest result for Brown Genres | Overall Accuracy of ML Algorithm |
|---|---|---|---|
| ZeroR | 0.13 | 0.00 | 6.93% |
| JRip | 0.14 | 0.24 | 11.38% |
| NNGE | 0.70 | 0.33 | 23.95% |
| Decision Table | 0.69 | 0.28 | 22.30% |
| REP Tree | 0.79 | 0.30 | 24.00% |
| J48 Tree | 0.83 | 0.35 | 24.84% |
| SMO | 0.85 | 0.44 | 30.94% |
| Nave Bayes | 0.83 | 0.30 | 23.98% |
| BayesNet | 0.85 | 0.41 | 28.34% |
| Lazy KStar | 0.80 | 0.36 | 22.05% |

Table I. Machine Learning Algorithm Accuracy Legislation And Brown Genres

Further indicators that legislation is different from the Brown genres in respect of its parts of speech and chunk characteristics came from a larger dataset extracted from the Brown Corpus and the Legislative Corpus. This dataset consisted of 31482 datapoints of which the legislative data constituted 3185 datapoints and the remainder from Brown genres. Using Weka, all features except parts of speech and chunk data were removed. Features not having discriminative power were also removed, leaving 43 features. Principal components analysis was utilised to represent features as independent orthogonal variables, leaving 36

features. Machine learning was carried out on this dataset with similar results as above.

Visualization of some of these principal components (see Figure 8), suggest that legislation can also be very different in its parts of speech and chunk characteristics to other English 'genres'. This complements the finding above that legislative readability metric and surface feature characteristics are different to 'normal' English. Further work is required to characterise the nature of these differences in detail and how they may be related to readability of legislation. They are suggestive that to the extent that 'plain English' has been achieved in legislation, (if it has) it has not resulted in 'normal English'.

The study we report above, has a number of limitations that fu- ture research might address. Only one jurisdiction is examined. The linguistic features examined are limited to readability metrics, surface characteristics, parts of speech and chunking data. The machine learning studies reported above show that other linguistic factors can be effective discriminators and also need to be explored in the legislative context
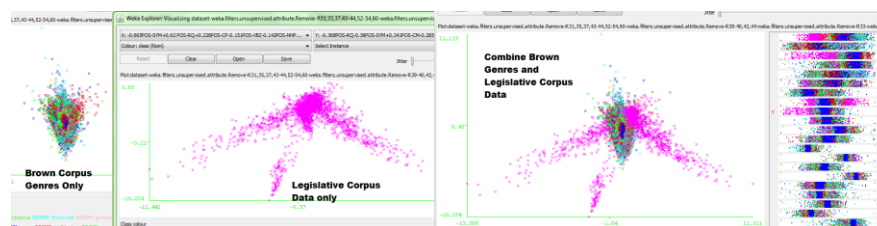


Fig. 8. Weka Visualizations of two principal components derived from parts of speech and chunk information (from left to right) for Brown Corpus Genres, Legislation Corpus and combined data

Every person who has read legislation knows that it is 'different'. What results such as the above show, is that it is possible to measure this difference. It is interesting that despite a commitment (and the considerable effort and expense in some cases) towards 'plain English' in

the drafting of laws, laws remain 'different' as a body of language (if we assume that the Australian Commonwealth legislative corpus is reasonably representative of legislative language in general). We are unaware of any past characterization of the empirical difference between a corpus of general English and a legislative corpus. An ability to define such points of difference, at a minimum can be envisaged to assist in identifying legislative sentences which are outside the umbrella of 'normal English usage'.

## 6. Conclusions and Future Work

This paper provides a background and context for carrying out readability research in application particularly to legislation with a particular focus on potential application of computational techniques. Empirical research on the readability of legislation supports a conclusion that most readers find it incomprehensible or difficult to read. Research on readability using natural language processing and machine learning is in its infancy, and is a promising area for further investigation. As far as we are aware there have not been significant studies on the readability of legislation applying crowdsourcing or machine learning techniques[12].

We report the development of the Readability Research Platform which is made available as an online service to researchers wishing to carry out readability research - whether on legislation (or other legal texts). We describe its envisaged use in a research context and report its performance characteristics.

Use of the Platform as a research tool is demonstrated in carrying out what is, as far as we are aware, novel empirical research assessing the difference between legislation and other written English using natural language processing and machine learning and examining readability metrics, surface features, parts of speech and chunk characteristics. Among our findings are that legislative data drawn from popular national

---

[12] Comparative corpora studies of legislation and other genres have previously been carried out in Dutch and Italian although not specifically in the context of readability issues.(van Noortwijk et al., 1995; Venturi, 2008).

legislation in one English speaking jurisdiction is different to 'normal' written English in respect of such characteristics at sentence level. Finding a difference is consistent with the empirical research which finds that legislative English is hard. How far we have come in achieving accessible legal language remains a live question. In addition, we undertake preliminary work on the use of parts of speech, chunk information, readability metrics and surface features to distinguish readability of sentences, using as input data, a corpus of graded reading material. This work shows such features to have discriminative value, but accuracy is low on a multiclass classification task. Readability metrics are, as others have observed, unreliable measures of readability, the more so in the context of legislation, given its difference from other English genres.

Finally, the establishment of the Readability Research Platform, we hope creates the potential (in combination with legislative sites and collaboration with other research groups) to carry out cloze testing and user evaluations on a large number of legal rules found in legislation. Such future studies, in our view, would be potentially make a valuable contribution to properly characterizing the readability of legislation. In particular, if a large dataset is created of legislative provisions labelled with reliable readability assessments, it can be expected to make available the full power of machine learning to identify those elements of legislative language which present a barrier to readability. At a minimum, it is likely to help us determine, with a greater level of confidence, how readable a particular piece of legislative text may be to its end users, without needing to undertake further human evaluations.

## 7. Appendix

These appendices provide examples of code used to run commands provided by the Readability Research Platform. Examples in Appendix A illustrate use of http requests to extract data. Appendix B provides python code to send multiple simultaneous commands and build a dataset for later machine learning.

# A. Simple http examples

## A.1. SINGLE COMMAND WITH SINGLE INPUT

This section illustrates sending a single command to the server using the iPython command line interface to send a command using python code. The output appears in blue. Line [1] imports the requests module which handles http requests. Line [2] defines the text to be analysed. Line [3] specifies which command is to be sent. Line [4] defines the url which is to be used (as described in the help page at the Readability Research Platform. Line [5] sends a http get request and saves the content to the variable 'output'. Line [6] prints the variable output to the screen. Lines [2]-[4] can be simplified to a single line but are expanded here to clarify the process.

```
Python 2.7.3 |Anaconda 1.4.0 (64-bit)

In [1]: import requests

In [2]: text = "The quick brown fox jumped over the lazy dog."

In [3]: command = "getallmetrics"

In [4]: url = 'http://buttle.anu.edu.au/readability/' + '?' +
                                    command + "XXXX" + text

In [5]: output = requests.get(url).content

In [6]: print request
```

fleschreadingease,fleschkincaidgradelevel,rix,colemanliau,

gunningfog,dalechall,ari,smog,lix::

103.70,1.03,0.00,4.43,3.60,0.45,6.62,3.00,9.00

## A.2. SIMPLE EXAMPLE USING TEXT FILE AND INPUT AND SAVING RESULTS TO OUTPUT FILE FOR LATER PROCESSING

The example below illustrates a simple use case where data analysis is carried out on an input text file. The results are saved to a file that can be opened in excel.

```python
# load python modules used in script

import requests

# open the text file to be used in read mode

textfile = open('demoparas.txt', 'r')

# split the document into a list of paragraphs

paragraphs = textfile.readlines()

# close the textfile - its not needed anymore

textfile.close()

# open a new datafile using .csv extension in write mode

# csv means a comma delimited file and can be read by excel

datafile = open('demoresults.csv','w')

# create an url & command variable

# ('?getari' and 'getfleshkincaidgradelevel' in this example)

url = ''http://buttle.anu.edu.au/readability/

commandurl1 = url + ''?getariXXXX"

commandurl2 = url + ''?getfleschkincaidgradelevelXXXX"

# loop through each paragraph and submit to

# the Readability Research Platform

# server, saving results to datafile

'∴n' inserts a line break after each data item

for para in paragraphs:

# get the results from each command
```

```python
result1 = requests.get(commandurl1 + para).content
result2 = requests.get(commandurl2 + para).content

# create a line to be written to the datafile

results = result1 + ',' + result2 + '∴n'

# print out to screen as well

print results

datafile.writelines(results)


# close the datafile

datafile.close()
```

## B. Example Script and Code for Data Extraction from the Readability Research Platform (http://buttle.anu.edu.au/readability/)

B.1. COMMANDS SENT USING IPYTHON TO RUN EXTRACTION SCRIPT AND THE WEKATOOL, WHICH SAVES DATA IN WEKA COMPLIANT FORMAT

The example below assumes that you have installed iPython, which makes running python code easier and comes with key libraries such as the Natural Language Toolkit already included. The text below is an illustration of the commandline interface in iPython with the two commands that would be needed to run the scripts and code in Appendix B.

```
Python 2.7.3 |Anaconda 1.4.0 (64-bit)

IPython 0.13.1 -- An enhanced Interactive Python.

[1] cd "D://YourDirectoryHoldingTheScripts/"
```

39

```
[2] run yourExtractionScript.py
```

## B.2. EXAMPLE EXTRACTION SCRIPT

The following is an example of a script run to extract data by sending multiple commands to the Readability Tool. The script is run from iPython as illustrated in Appendix B.1. Copy and save the script with an appropriate name - 'yourExtractionScript.py'. In the following code, comments describing the code are in dark green and are not executed by the computer.

```python
# load code for holding/processing data as Weka format

import wekatool as weka

import os, nltk

# The list of data commands to be sent to the server

commandList = [['getallmetrics'],['getsurfaceD','normed']]

commands = str(commandList)

#output file where results will be saved

outputfile = 'legislation1.arff'

# Load the wekaTool for later use

wkT = weka.wekaTool()

# Change to directory of your legislation corpus

os.chdir('D://PhD/A-Local/yourLegislationCorpus/')

# get the names of text files to be processed

filelist = []

for file in os.listdir("."):

    if file.endswith(".txt"):
```

40

```python
        filelist.append(file)

# for each text file process the file

for file in filelist:

    # provide feedback on progress

    print "STARTING ON FILE: ", file

    #assign a class to data as required

    classType = 'legislation'

    f = open(file).read()

    # splitting the file into sentences

    sentences = nltk.sent_tokenize(f)

    count = 1

    #For each sentence in the file process the sentence

    for sentence in sentences:

        print "PROCESSING SENTENCE: ", count

        count +=1

        # run the weka tool to load

        # the data item for later processing

        wkT.loadTextData(sentence,commands,classType)

# process the data and write it to file

#  for  later  use  for  machine  learning
arff = wkT.writeARFFfile(outputfile)
```

## B.3. EXAMPLE PYTHON CODE FOR EXTRACTING DATA IN WEKA FORMAT

The following code can be used with the commands illustrated in Appendix B.1 and the script example in Appendix B.2. The entire code below can be saved into a file called 'wekaTool.py', after which can be called by code illustrated above.

```python
from __future__ import division

import requests, urllib2, math, re, traceback, sys, ast

"""

A class for extraction of features from text.

This code is developed as part of PhD studies in the
ANU Research School of Computer Science.

It may be freely used for research purposes only.
For other uses, contact the author.

Author: Michael Curtotti 2013

"""

class wekaTool:

    """

    command syntax:

    [[command,mode,engine,type,ngramcount],[...],[...],...]

    the first value is required

    the 2nd to 4th values are optional

    mode = raw or normed

    engine = monty or nltk
```

```python
type = letter or word or pos

test command = ['getposd','normed'],

['getchunkd','ra
w','monty'],

['getsurfaceD','ra
w','monty'],

['getngram','raw','monty','le
tter','1,2'],

['getngram','raw','monty','
pos','1,2,3']

# holds data for a single input after which it is cleared

featureDictionary = {}

# a holder for keys for features across many data items

featureList = []

# a holder for data extracted from text input

# holds multiple inputs for later data formatting

# inputs for each text item will be held as python

# dictionary objects with each key representing

# a feature and each value the value of that feature.

dataset = []

url = ""

errorCount = 0 inputCount = 0

def                                                   __init__(self,
url='http://buttle.anu.edu.au/readability/'):
```

```python
    """
    A class for creating a feature set from text. Supply
    the url for code testing purposes only
    """
    self.url = url
    self.errorCount = 0 self.inputCount = 0

def loadFile(self, text = "", commands = [],
classType = "UNK"):
    """
    loads an entire file, partitioning the input into
    sentences
    Used as alternative to the loadTextData function
    Needs ['partition'] to be included in
    list of commands
    """
    try:
        commands = str(commands)
        body                                    =
        {'commands':[commands],'text':[text],'class':[cl
        assType]}
        result = requests.post(self.url,body).content
        assert    not    result.startswith('ERROR')
        processedresult = ast.literal_eval(result)
    self.dataset += processedresult
    self.inputCount += 1
except Exception, e:
```

```python
        self.errorCount += 1

        print "ERROR with input"

        print "Number of Errors: ", self.errorCount

        print "Number    of    Successful    inputs:    ",
        self.inputCount print "TEXT WAS: ", text[:200]

        print "COMMANDS WERE: ", commands

        print traceback.print_exc()

        _,_,tb = sys.exc_info()

        traceback.print_tb(tb)

        print "===================="

def loadTextData(self,text = "", commands = [], classType =
"UNK"):

    """

    processes text data by calling the Readability Tool

    at   http://buttle.anu.edu.au/readability/   receives   data
    extracted from the input text and holds it

    for later output to file or
    printing

    """

    try:

        commands = str(commands)

        body                                              =
        {'commands':[commands],'text':[text],'class':[classTy
        pe]}

        result = requests.post(self.url,body).content
```

```python
            assert not result.startswith('ERROR')

            assert len(body.keys())>0

            processedresult = ast.literal_eval(result)

            if not len(processedresult.keys())==0:

                self.dataset.append(processedresult)

            self.inputCount += 1

        except Exception, e:

            self.errorCount += 1

            print "ERROR with input"

            print "Number  of  Errors:  ",
            self.errorCount

            print "Number  of  Successful  inputs:  ",
            self.inputCount

            print "TEXT WAS: ", text

            print "COMMANDS WERE: ", commands

            print traceback.print_exc()

            _,_,tb = sys.exc_info()

            traceback.print_tb(tb)

            print "===================="

def __buildFeatureList__(self):
    """

    internal method for building a list of all features.

    """
```

```python
        for item in self.dataset:

            for key in item.keys():

                if not key in self.featureList:

                    self.featureList.append(key)

    def writeARFFfile(self,filename='data.arff'):

        """

        writes ARFF data to file

        Do not run until all data has been generated

        Using the loadTextData method or the loadFile method

        """

        data = self.createARFF()

        arfffile = open(filename,'w')

        arfffile.writelines(data)

        arfffile.close()

    def createARFF(self):

        """

        returns a arff format string

        Do not run until all data has been generated

        This is intended as an internal method

        use createARFF method instead"""

        self.__buildFeatureList__()

        string = self.getArffHeader()
```

```python
        string += "@data\n"

        count = 1

        for item in self.dataset:

            count +=1

            string += self.getArffItem(item, 'arffsparse')

        return string

def getArffHeader(self):

    """

    returns string for arff header -

    do not run until all data has been generated

    internal method for ARFF data generation

    """

    string = "@RELATION dataset\n"

    string += "\n\n"

    #string +='@ATTRIBUTE dummystring STRING\n'

    classtypes = []

    for item in self.featureList:

        if item == 'inputText':

            string += "@ATTRIBUTE " + item + ' ' + 'STRING\n'

        elif item != 'classType':

            item = item.replace(',','CM')

            item = item.replace('"','LDQ')
```

48

```python
            item = item.replace("'",'LQ')

            string += "@ATTRIBUTE " + item.replace(',','CM') +
            ' ' + 'NUMERIC\n'

    for item in self.dataset:

        if not item['classType'] in classtypes:

            classtypes.append(item['classType'])

    string += "@ATTRIBUTE class {"

    for item in classtypes:

        string += item + ","

    string = string[:-1] +"}"

    string += '\n\n'

    return string

def getArffItem(self, fdict = {}, format='arffsparse'):

    """ internal method for generating an individual weka
    format data feature set from loaded data - do not run until
    data is loaded

    """

    string =""

    if format == 'arffsparse':

        string +="{" #string +="{0 'dummyvalue',"

    tuples = []

    ARFFfeatureList = []

    ARFFfeatureList = self.featureList if 'classType' in
    ARFFfeatureList:
```

```python
        ARFFfeatureList.remove('classType')

for key in fdict.keys():

    if not key == 'classType':

            # we use the key to get

            # the index number for the data point

            index = ARFFfeatureList.index(key)

            # we create a tuple from the index,

            datapoint = str(fdict[key])

            datapoint = datapoint.replace(',',' CM')

            datapoint = datapoint.replace('"',' DQ ')

            datapoint= datapoint.replace("'",' SQ ')

            tuples.append(((index),datapoint,key))

            #print index, fdict[key].replace(',','CM'), key

tuples.sort()

for tup in tuples:

    if tup[2] == 'inputText':

            string += str(tup[0])+ ' "' + tup[1] + '",'

    elif not tup[2] == 'classType':

            if not float(tup[1]) == 0:

                string += str(tup[0])+ ' ' + tup[1] + ','

string       +=       str(len(ARFFfeatureList))       +       '
"'+fdict['classType']+'"'
```

```
    string +="}\n"

elif format == 'arff':

    pass

return string
```

## Acknowledgements

Our research would not have been possible without access to soft- ware packages made freely available by other researchers and individuals in particular the NLTK Natural Language Toolkit, the Weka Machine Learning Package and Montylingua. We gratefully acknowledge the work of these researchers. We also thank the reviewers whose suggestions have assisted us in improving this final version of the paper.

## References

Eloise Abrahams (2003), *Efficacy of plain language drafting in labour legislation*. Master's thesis  on Human Resource Management), Cape Peninsula University of Technology, South Africa.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton (2010), *Readability assessment for text simplification*. In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp. 1-9.

F.A.R. Bennion (1983), *Statute law*. Oyez

Robert W. Benson (1984), *End of legalese: The game is over.* NYU Review of Law & Social Change , Vol. 13, p. 519.

Steven Bird, Edward Loper, and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

J.R. Bormuth (1967), *Cloze readability procedure*. University of California Los Angeles.

Kevyn Collins-Thompson and James P Callan (2004), *A language modeling approach to predicting reading difficulty.* In HLT-NAACL, pp. 193-200.

O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken (2013), *Using the crowd for readability prediction*. Natural Language Engineering, pp. 1-33.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi (2011), *Read-it: Assessing readability of Italian texts with a view to text simplification.* In Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, pp. 73-83.

W.H. DuBay (2004), *The principles of readability.* Impact Information, pp. 1-76.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad (2010), *A comparison of features for automatic readability assessment.* In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, pp. 276-284.

W.N. Francis and H. Kucera (1964), *A Standard Corpus of Present-Day Edited American*. Revised 1971, Revised and Amplified 1979. Department of Linguistics, Brown University Providence, Rhode Island, USA. Available at: www.hit.uib.no/icame/brown/bcm.html (accessed 10th December, 2013)

GLPi and V. Smolenka (2000), *A Report on the Results of Usability Testing Research on Plain Language* Draft Sections of the Employment Insurance Act. Available at:
http://www.davidberman.com/wp-content/uploads/glpi-english.pdf
(accessed 10th December, 2013)

Mark Hall, Eibe Frank, Geofrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). *The weka data mining software.* ACM SIGKDD Explorations, Vol. 11, No. 1.

J. Harrison and M. McLaren (1999), *A plain language study: Do New Zealand consumers get a "fair go" with regard to accessible consumer legislation.* Issues in Writing, Vol. 9, pp. 139-184.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi (2008), *An analysis of statistical models and features for reading difficulty prediction.* In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp. 71-79.

P. Heydari and A.M. Riazi ( 2012). *Readability of texts: Human evaluation versus computer index.* Mediterranean Journal of Social Sciences, Vol. 3 No. 1, 2012, pp. 177-190.

Miller J. (2005), *The development of the legal information institutes around the world.* Canandian Law Library Review, Vol. 30, No. 1, p. 8

Simon James and Ian Wallschutzky (1997), *Tax law improvement in Australia and the UK: the need for a strategy for simplification.* Fiscal Studies, Vol. 18 No. 4, pp. 445-460

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty (2010). *Learning to predict readability using diverse linguistic features.* In Proceedings of the 23rd International Conference on Computa- tional Linguistics, Association for Computational Linguistics, pp. 546-554

J. Kimble (1994), *Answering the critics of plain language.* The Scribes Journal of Legal Writing, Vol. 5, p. 51.

G.R. Klare (2000), *Readable computer documentation.* ACM Journal of Computer Documentation (JCD), Vol. 24, No. 3, pp. 148-168

Uta Kohl (2005), *Ignorance is no defense, but is inaccessibility? On the accessibility of national laws to foreign online publishers.* Information & Communications Technology Law, Vol. 14, No. 1, pp. 25-41

Hugo Liu (2004), *Montylingua: An end-to-end natural language processor with common sense.* Available at: http://web.media.mit.edu/~hugo/montylingua/ (accessed 10th December, 2013)

P.W. Martin (2000), *The mushrooming virtual law library on the net*. In Cornell Law Forum, Vol. 27.

D. Melham (1993), *Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth*. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs.

Jay Milbrandt and Mark Reinhardt (2012), Access Denied: Does Withholding the Law Violate Human Rights? Regent Journal of International Law, *Forthcoming.* Available at SSRN: http://ssrn.com/abstract=2132672 (accessed 10th December, 2013)

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily (2010), *Crowdsourcing and language studies: the new generation of linguistic data.* In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, pp. 122-130.

PCO NZ (2007), *Presentation of New Zealand Statute Law*: Issues Paper 2. Technical Report 2, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office.

PCO NZ (2008), *Presentation of New Zealand Statute Law*. Technical Report 104, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office.

OLR (2003), *Inland Revenue Evaluation of the Capital Allowances Act 2001* rewrite, Opinion Leader Research. Technical report, UK Inland Revenue.

OPC-Australia (2003), *Plain English*. Technical report, Australian Common wealth Office of Parliamentary Counsel.

OPC-UK (2013), *When Laws Become Too Complex: A Review into the Causes of Complex Legislation.* Technical report, United Kingdom Office of Parliamentary Counsel.

PCO-NZ (2011). *A Review of Methods for Measuring the Quality of Legislation.* Technical report, New Zealand Parliamentary Counsel's Office.

N. Pettigrew, S. Hall, and D. Craig (2006), *The Income Tax (Earnings and Pensions) Act* - Post-Implementation Review, Final Report MORI.

Emily Pitler and Ani Nenkova (2008), *Revisiting readability: A unified framework for predicting text quality.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 186-195.

G. Richardson and D. Smith (2002). *Readability of Australia's goods and services tax legislation: An empirical investigation*, Federal Law Review, Vol. 30, p. 475.

Adrian Sawyer (2010), *Enhancing compliance through improved readability: Evidence from New Zealand rewrite experiment*. Recent Research on Tax Administration and Compliance.

Sarah E Schwarm and Mari Ostendorf (2005), *Reading level assessment using support vector machines and statistical language models.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 523-530

Luo Si and Jamie Callan (2001), *A statistical model for scientific readability.* In Proceedings of the tenth international conference on Information and knowledge management, ACM, pp. 574-576.

Johan Sjöholm (2012), *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. PhD thesis, Linköpings University, Sweden. Available at: http://www.ida.liu.se/projects/webblattlast/Rapporter/lasbarhet.pdf (accessed 10th December, 2013)

D. Smith and G. Richardson (1999), *The readability of Australia's taxation laws and supplementary materials: an empirical investigation.* Fiscal Studies, Vol. 20, No. 3, pp. 321-349.

Edwin Tanner (2002), *Seventeen years on: Is Victorian legislation less grammatically complicated.* Monash University Law Review, Vol. 28, p. 403.

C van Noortwijk, RV De Mulder, and RW van Kralingen (1995), *Word use in legal texts: statistical facts and practical applicability.* Legal Knowedge Based Systems: Telecommunication and AI & Law (JURIX95), Lelystad: Koninklijke Vermande, pp. 91-100.

G. Venturi (2008), *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, p. 1.

G. Wagner (1986), *Interpreting cloze scores in the assessment of text readability and reading comprehension.*

B. Woods, G. Moscardo, T. Greenwood, et al. (1998), *A critical review of readability and comprehensibility tests.* Journal of Tourism Studies, Vol. 9, No. 2, pp. 49-61

# Software Tools for the Visualization of Definition Networks in Legal Contracts

## Research Abstract[*]

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

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

Srinivas Sridharan
University of California
San Diego
California, United States
srsridharan@ucsd.edu

## ABSTRACT

This paper describes the development of prototype software-based tools for visualizing definitions within legal contracts. The tools demonstrate visualization techniques for enhancing the readability and comprehension of definitions and their associated characteristics. This contributes to more accurate and efficient drafting or reading of contracts through the exploration of the meaning and use of definitions including via word clouds, multilayer navigation, adjacency matrix and graph tree representations.

## Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Natural language; H.5.4 [**Hypertext and Hypermedia**]: Navigation; I.7.2 [**Document Preparation**]: Format and Notation; I.7.5 [**Document Capture**]: Document Analysis

## General Terms

Human Factors

## Keywords

definitions, legal contracts, word clouds, network visualization, contract visualization, text visualization, graph metrics

## 1. INTRODUCTION

This paper addresses the visualization of definition use within contracts. It is part of ongoing research on the development of software-based tools for reading and writing legal rules in contracts and legislation and aims to improve accessibility of legal documents and increase the efficiency and accuracy of legal rule creation [4, 5, 6].

This paper reports the development of prototype software tools demonstrating novel applications of visualizations for the representation and analysis of definition networks within contracts. The

software tool enables a user to input text via a web interface and presents the user with a number of alternative visualizations of definitions in a contract: single layer pop-up hyper-linking of defined terms as they are used and representation of frequency and other information; application of 'word cloud' techniques to enable the rapid and global visualization of the 'usage' of a defined term and 'obfuscation' of a defined terms (metrics reflecting both the semantic content and graph theoretic role of the term); multi-layer hierarchical navigation tools enabling in-situ navigation of 'definition networks' from the rule where a definition is used; visual presentations of definitions as a link and node graph; and matrix representation of definition usage within a contract.[1] In Section 6 below we describe these visualizations further.
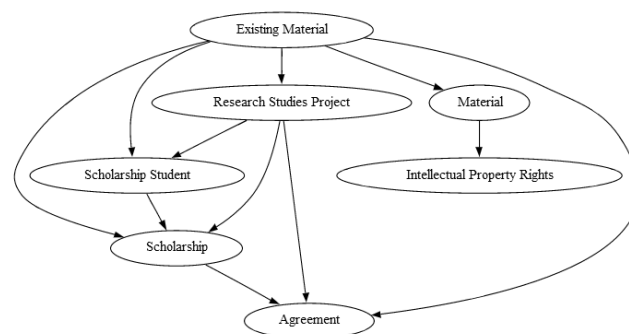


**Figure 1: A node-link graph diagram showing the relationships between defined terms which have been extracted from a natural language contract.**

Contracts are semi-structured documents, and are usually explicitly organized in a tree-like structure consisting (primarily) of rules and sub-rules. In the Australian case these structures are typically referred to as 'clauses' and 'sub-clauses' with each clause ideally addressing a discrete topic. Definitions typically occur as a small glossary or dictionary embedded within a single 'definition' clause.

Definitions form a substantial part of typical contract texts and are used by drafters to control meaning and presentation of text. In this paper we use 'defined term' to refer to the definition label and 'defining text' to refer to the natural language which expresses the

[1]http://buttle.anu.edu.au/contracts/

meaning of the defined term. 'Definition' refers to the entire structure. While formally definitions are intended to simplify drafting, they can also be used in larger contracts as a tool to modify meaning in the favour of the drafter's client in ways that become increasingly difficult to analyze for the other party as the complexity of definitional relationships increases. Such definitions can also result in meaning being 'hidden', as meaning may not be apparent from the surface text of a legal rule. Because of their complexity, such structures can also result in errors such as inconsistency of meaning in a hierarchy of definitions.

This paper is organized as follows. Section 2 reviews relevant literature. Section 3 describes the extraction of definitions from a contract document. Section 4 discusses the representation of definitions and their relationships as networks. Section 5 briefly outlines the data and tools used in undertaking this work and briefly canvasses analysis of that data. Section 6 presents prototype visualizations exploiting the network characteristics associated with definitions. We present our conclusions in Section 7.

## 2. RELATED WORK
There are four areas of related work we wish to describe: work relating to the study of contracts at the broadest level; natural language processing for the extraction of definitions; information and graph visualization and studies in relation to Word Clouds.

### 2.1 Studying Contracts
Contracts are studied from a wide range of perspectives and disciplines. The principles for interpreting contracts as sources of legal rules is an extensively studied domain. Contracts have also been widely studied from the point of view of economic and social theory [19]. Work more directly relevant to the high level aim of creating software-based tools to enhance the reading and writing of contracts is also potentially broad. Curtotti et al. [4, 5] review work including in the field of machine learning, the logical representation of legal rules, e-contracts and studies of corpora of contracts.

### 2.2 Natural language processing for the extraction of definitions
Work on the application of natural language processing to definitions in general text is extensive, however a considerable part of this work is dedicated to extraction of definitions from unstructured general prose. It thus addresses a more complex and difficult problem than that of extraction of definitions from semi-structured texts, such as contracts. Degorski et al. [9] apply enhancements to machine learning for definition extraction from unstructured text. Others employ rule based approaches for the extraction of dictionaries from text [17]. Winkels et al. [20] and Maat et al. [7] report work in the parallel legislative domain including definition extraction, in the context of Dutch legislation. Definition extraction remains an active area of research with a view to improving precision and recall of such definition extraction [18].

### 2.3 Information and Graph Visualization
Information visualization is centred on the users of data and is concerned with the representation of complex data in ways that facilitate its comprehension. Information visualization employs graphical presentations of data to exploit the visual capacities of users in identifying patterns and relations in data. It is used in text mining and may provide advantages such as the ability to display a large amount of data at once, enhance identification of relationships and clustering in data, provide interactivity to users or allow users to move from micro to macro quickly [10, pp 190 et seq]. Some forms of such data visualization are commonly known (e.g. histograms and line graphs). Others are more recent technologies developed for the visualization of large data sets. Concept set graphs are a commonly used tool in text mining showing hierarchical relationships between concepts. Graphs may also show the network of associations between concepts or entities found in texts, and the weight of those associations. Circle graphs can be used to show the strength of multiple associations between terms. A plethora of more complex visualizations have also been employed including self organising maps, hyperbolic trees and fisheye diagrams[10, pp 194 et seq]. Among the variables that can be adjusted to enhance graph visualization are layout (including tree layout, 3D representation, spring layout, space division and matrix layout), clustering, sampling or filtering for large graphs, zooming and panning, animation, focus plus context [3]. A site that provides both a software tool for a range of common visualizations and demonstrations of their application is the Java Infovis toolkit site.[2] Visualizations are commonly employed in the field of network analysis (including for example analysis of social networks). Most commonly as the node and link diagram used in graph theory, but enhanced with information describing the entities and relationships represented by nodes and links. An alternative representation also employed in social network analysis are adjacency matrices which provide a two dimensional array representing nodes and associations (or strength of association) between them [pp 4 et seq and pp 259 et seq][8].

### 2.4 Word Clouds
Word clouds are a form of information visualization that has become popular in recent years as a way of summarising and visualizing key concepts in a large body of text. Word clouds support functions such as browsing, searching, subject description and formation of an impression concerning the data. A key technique in word clouds is the manipulation of the visual features of text (font, area, width, intensity, colour) and their location within the cloud to suggest importance or other features of the word. Bateman et al. [1] find that font size and weight has a particular effect. Colour can also influence interaction but is ambiguous in its meaning. Position also has an influence. Accordingly they endorse the use of the former while suggesting that colour and position be used with care. Lohmann et al. [16] specifically study the effect of tag position or layout on the effectiveness of certain user tasks such as identification of popular terms, search for particular terms and identification of topics in the word cloud. Based on studies of user interaction with different layouts they do not find a best way to layout a cloud but observe that large tags (font size) are readily identified as 'popular'. They confirm findings by other authors that centering of 'popular' tags within a cloud assists their identification. This effect they find most pronounced with a circular tag layout. They also find the top left quadrant of a word cloud attracts the most attention. Word clouds are not well suited for searching. Halvey et al. [14] also find that font size and position are important, although they note that alphabetical presentation is an aide to finding information.[3]

---

[2]thejit.org

[3]We note that there is work in the parallel domain of visualization of legislation. Due to limitations of space we do not canvas that work here, but refer the interested reader to *The Visualization of Law*, Curtotti and McCreath [6]. Also work on visualizing contract provisions using non-computerized methods has been undertaken by Haapio and Passera [11, 12].

**DEFINITION CLOUD**

Agreement • Commercialise • Confidential Information • Entry Requirements • Existing Material • GST • GST Law • Intellectual Property Rights • Material • New Material • Parties • Principal Supervisor • Research Studies Program • Research Studies Project • Scholarship • Scholarship Amount • Scholarship Student • Supply • Tax Invoice • Thesis

**CONTRACT WORD CLOUD**

existing program presentation sent confidential liability facsimile new licence address prior addressee rights tax information provide law acknowledges consent including thesis material day clauses research receiving given days student clause number publication b accordance person performance shall mediation request scholarship agreed relation principal dispute means subject project written use lug notice extent received c s respect writing property contract pay party reference gst studies costs schedules loss does provider payments schedule agreement limitation intellectual right parties supervisor time includes enrolled

**Figure 2: Definition cloud and comparative word cloud.**

## 3. EXTRACTION OF DEFINITIONS FROM CONTRACTS

To visualize definitions in a contract it is necessary to first extract them and clauses from the contract. We use relatively trivial regular expressions which are applied in three stages: (1) identification and segmentation of the definition clause and other clauses in the text of the contract; (2) segmentation of definitions from each other; and (3) extraction of the defined term and its defining text from a definition. By requiring users to apply simple rules which are widely used in Australian industry practice, such as ensuring definitions ends with a full stop and using standard 'key words' for definition relations (particularly the words 'means' and 'includes') essentially 100% accuracy can be attained on typical contract texts. This result expresses that a realistic and readily attainable solution (minor user editing) can effectively address the accuracy problem which is difficult to fully solve using entirely computational methods. Changes necessary to improve accuracy are easily exposed to the user through a web page and are implementable with a few key strokes. Previous research by the authors using fully automated methods of machine learning, hand crafted rules or hybrid methods reached accuracies of around 80% to 82%, a level of accuracy inadequate to the legal domain [4].

## 4. REPRESENTATION AS NETWORKS

A graph is an ordered tuple $G = (V, E)$ of a set of vertices (or nodes) 'V' and edges 'E' between them. An edge links two vertices $v_1$ and $v_2$, and may either be directed or undirected [2, p348]. The number of edges associated with a vertex is referred to as its 'degree'. In the case of a directed graph, the in-degree of a vertex is the number of incoming arcs to the vertex. Its out-degree is the number of arcs emerging from the vertex [2, p348 et seq]. We follow de Nooy et al. in defining a 'network' as a graph which has additional information associated with its vertices and edges, i.e. information beyond the simple structural characteristics of nodes and links [8, p7]. Definitions and the clauses in which they occur are thus represented as the vertices (or nodes) of a network. Links between the vertices represent either the occurrence of a defined term in a clause, or the occurrence of a defined term in another definition. The nodes and links form directed graphs which can be analysed, including from a graph theoretic viewpoint to reveal information about a legal document. Figure 1 shows an example of a network between definitions in a contracts.

## 5. DATA, TOOLS AND ANALYSIS

The work reported in this paper is based on analysis of a set of ten contracts drawn from a corpus of 249 Australian contracts consisting of in the order of $10^6$ words. The corpus has been compiled from Australian contracts and contract drafts available on the web. Curtotti et al. [5] report the profiling and analysis of an earlier version of the corpus. The current version of the corpus has been subjected to further data cleaning but is substantively the same as reported above.[4]

The tools used in carrying out the work reported here included project code for analysis and visualization (including online) which is written primarily in python and javascript and libraries for graph analysis or graph visualization (networkX and Graphviz). We also utilize Canviz for web visualization of graphviz output.[5]

On average, definitions represent 17.4% of the core text of contracts in the sub-corpus and these contracts on average used definitions 304 times. These latter results illustrate the significance of definition text as a component of such legal documents.

Definition networks in our sub-corpus had on average a degree of 2.35 with a standard deviation of 2.47. This relationship between average and standard deviation would lead us to anticipate that degree is log-normally distributed [15]. This is in fact a reasonable description of the distribution.

Length is often used as a simple measure of complexity: the longer a definition, the more complex it is likely to be [13]. We found the correlation between out-degree and definition length to be low to moderate with a value of 0.325 over 223 data points (individual definitions). Degree then provides a different indicator of complexity to length.

In some of the visualizations described below we employ a recursive out-degree related measure to represent 'hidden' meaning in a definition. This measure is derived from the overall length of all text recursively referenced through the outward links of a defini-

---

[4]A copy of the corpus is obtainable by contacting the authors via http://cs.anu.edu.au/people/Michael. Curtotti/
[5]http://code.google.com/p/canviz/

tion. This is effectively a recursive weighted out-degree measure, where length is a measure of weight of the parent and successor nodes.

# 6. VISUALIZATIONS

We report a prototype website demonstrating a web based tool for the extraction and visualization of definition structures from submitted contract texts.[6] A user is able to load the demonstration text or submit a contract conforming to the requirements of the tool and may visualize definition structures within the contract by selecting one of four visualization options.

**Cloud Visualisations:** Figure 2 illustrates a cloud presentation of two measures of definition characteristics: their frequency of use in a contract (usage) (visualized through font size), and how much of the meaning of a defined term is 'hidden' or 'obfuscated' through the referencing of other defined terms by a defining text. Red, yellow, green is a well recognized 'traffic light' representation suggestive of levels of risk and is used in this context to indicate risk prone definition relationships with red suggesting significant hiding of meaning, yellow moderate hiding and green a low proportion of hidden text.

The scaling of font size makes it relatively straight forward to determine the probable purpose of the document from which the definitions are drawn as significant terms are emphasised. A more traditional word cloud is provided from the same contract for the purposes of comparison. In this word cloud word font size is a function of word frequency.

Use Case: Such a visualization allows a reader to form an immediate impression of the importance of terms, where complex layered meaning may be hidden and the probable nature of the contract.

**In Situ Usage and Obfuscation:** Similar information to that conveyed via the Definition Cloud is provided by in situ presentations using a number (to directly represent usage) and a small pie chart icon to represent 'obfuscation'. In this case the ratio of the pie shown in red represents the relative length of hidden text associated with the definition. (See Figure 3)

Use Case: As above.
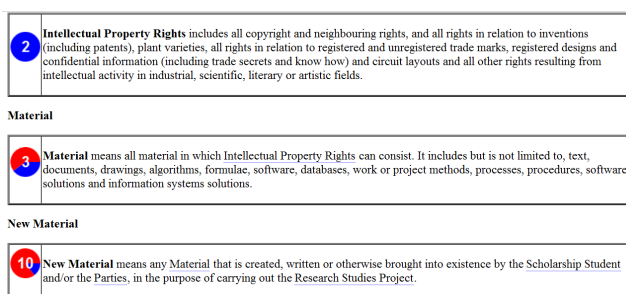


**Figure 3: Usage and Obfuscation metrics**

**Single and Multilayer in-situ Definition Graph Navigation:** In most contracts, the only means of navigating to the meaning of a defined term when encountering it in a clause is to scroll to the

---

[6] http://buttle.anu.edu.au/contracts/

definition clause (typically at the top of the document), read the definition and scroll back down.

We provide both a visualization that allows single layer access to the meaning of a defined term and navigation through multiple layers of definition referencing. The latter visualization enables a user, from the rule being read, to navigate through the entire tree of definitions referenced by the rule, following the conceptual links between defined terms. By double clicking the definition window the user can cause the pop up to disappear, returning to the original rule.

Use Case: Such a facility is likely to aid both comprehension and increase efficiency of contract reading. It reduces time necessary to access related meanings and anchors the reading experience in the rule itself.
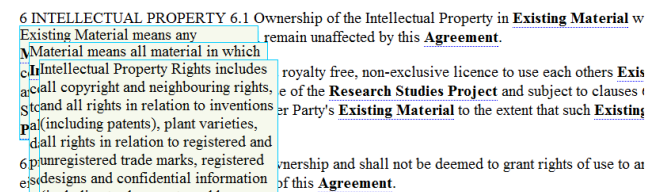


**Figure 4: Tool for multi-layer navigation of defined terms from rule where the defined term is used. Here we navigate the terms 'Existing Material' – 'Material' – 'Intellectual Property Rights' from the clause of the contract dealing with those rights.**

**Matrix Representation of Bimodal Definition Use Graph:** Figure 5 provides a representation of the relationship between clauses and definitions as a weighted bimodal adjacency matrix. Such representations are used in social network analysis,[8] but their application to legal documents is novel.

Each square in the matrix represents a definition-clause relationship and the darkness of the square indicates the relative frequency with which a defined term is used in a particular clause. A column provides a visual summary of the importance of definitions used within a particular clause, while a row summarises the use of a particular definition across the agreement.

Use Case: The bimodal representation provides a potential tool for visualizing the semantic structure of a contract in summary form.

**Definition Graphs:** Figure 1 is an example of standard node link graph diagram representing a definition network. It shows the relationships between a definition and the defined terms it uses. The visualization provides an immediate sense of the relationship between defined terms. It intuitively represents the complexity of definition use, providing an opportunity to a drafter to consider revision to reduce complexity, or to a reader to explore concepts utilised by a rule. A reader is similarly alerted to semantic relationships. Simple inspection reveals any cycles that may be present in the definition graph. Cycles may represent logical errors or conceptual complexity in the ideas represented by the definition. Graphs of this kind can equally be generated with a rule as the root node of the representation. Although the adjacency matrix visualization provides an indication of 'weight', it only indicates relationships of a clause or rule with definitions to a depth of 1 (i.e. those directly
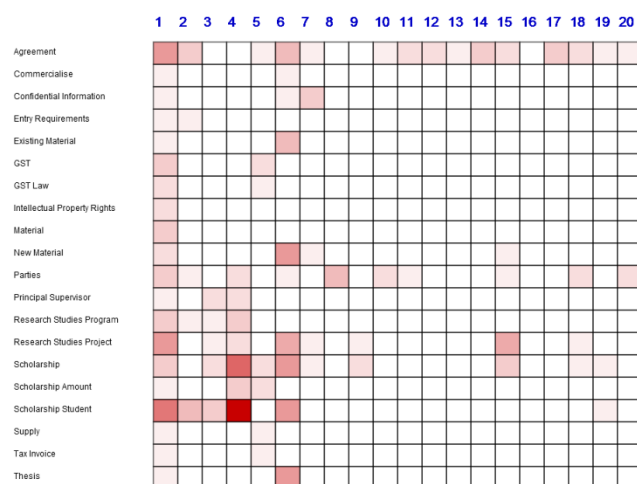
**Figure 5: A matrix representation of the relationships between definitions and clauses. Extracted using Pajek.**

used in the clause text). A directed node link diagram enables the relevant definition network to be explored in full.

Use Case: Provides a graphical representation of the semantic structure of key terms in a contract, assisting readers in understanding semantic relationships and drafters in removing potential errors or simplifying how terms are defined.

## 7.  CONCLUSIONS AND FUTURE WORK

In this paper we present work related to the visualization of definition networks. We describe definition usage in contracts and present a number of prototype visualizations of definitions (including visualization of network attributes and selected metrics). Methods widely employed outside the legal field (such as word clouds) show promise for application within the legal field in connection with facilitating comprehension of definition use in contracts. Navigational enhancements such as multi-layer pop up for definition navigation show the potential to facilitate access to the meaning of definitions within the context of rules in which they are employed increasing comprehension and efficiency in contract reading. Tools such as node-link diagrams facilitate an exploration of semantic trees embedded in definition networks. Presentation of metrics associated with the definition network help readers assess the significance and risk of defined term usage.

## 8.  REFERENCES

[1] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202. ACM, 2008.

[2] I. Bronstein, K. Semendyayev, G. Musiol, and H. Muehlig. *Handbook of Mathematics Fifth Edition*. Springer, 2007.

[3] W. Cui. A survey on graph visualization. *PhD Qualifying Exam (PQE) Report, Computer Science Department, Hong Kong University of Science and Technology, Kowloon, Hong Kong*, 2007.

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

[5] M. Curtotti and E. McCreath. A corpus of australian contract language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law 2011*, 2011.

[6] M. Curtotti and E. McCreath. Enhancing the visualization of law. In *Law via the Internet Twentieth Anniversary Conference, Cornell University*, 2012.

[7] E. de Maat and R. Winkels. Automatic classification of sentences in dutch laws. In *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 207–216. IOS Press, 2008.

[8] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.

[9] Ł. Degórski, M. Marcinczuk, and A. Przepiórkowski. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, 2008.

[10] R. Feldman and J. Sanger. *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[11] H. Haapio. Contract clarity through visualization–preliminary observations and experiments. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 337–342. IEEE, 2011.

[12] H. Haapio and S. Passera. Reducing contract complexity through visualization-a multi-level challenge. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 370–375. IEEE, 2012.

[13] J. Hagedoorn and G. Hesen. Contractual complexity and the cognitive load of r&d alliance contracts. *Journal of empirical legal studies*, 6(4):818–847, 2009.

[14] M. Halvey and M. Keane. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314. ACM, 2007.

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

[16] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. *Human-Computer Interaction–INTERACT 2009*, pages 392–404, 2009.

[17] S. Muresan and J. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference*, volume 1, page 30, 2002.

[18] R. Navigli and P. Velardi. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327. Association for Computational Linguistics, 2010.

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

[20] R. Winkels and R. Hoekstra. Automatic extraction of legal concepts and definitions. In *Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference*, volume 250, page 157. IOS Press, 2013.

# Enhancing the Visualization of Law

Michael Curtotti

Research School of Computer Science

Australian National University

michael.curtotti@anu.edu.au

Eric McCreath

Research School of Computer Science

Australian National University

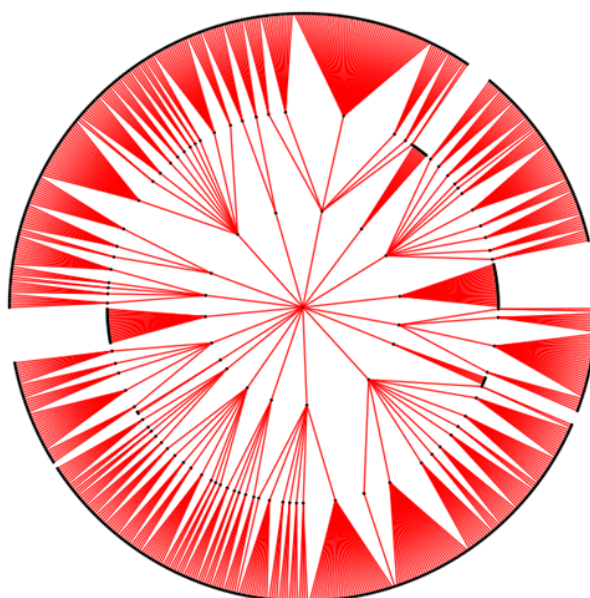eric.mcreath@anu.edu.au *

October 2, 2012



**Figure 1:** The formal structure of the Copyright Act 1968 (Aust)

## Contents

1

# 1   Introduction

The accessibility of law has undergone a revolution in the last two decades as public good, official and commercial initiatives have made legislation (and other legal materials) accessible onlinet. In respect of legislation, this development has followed centuries of refinement in how the law is written and presented. The presentation of the law (or in its 21st century manifestation - its visualization) has long been known to influence its readability (itself a dimension of the accessibility of law).

Online legislation sites vary widely in their approaches. The most basic present legislation as a scrollable text (in practice reverting to the equivalent of a single scroll of paper), the most novel use features such as colour, graphs, images, moving pictures and information enhancement to improve visualizations. Some sites focus on providing laws as downloadable documents in various formats – emphasising the online availability of 'the official version'. Many sites provide access or links to accompanying materials such as explanatory memoranda, subordinate legislation or court interpretations. Some sites offer legal rules within legislation as navigable nodes, providing links to key information including (in some cases) links to cross references and defined terms used in a legal rule. Search tools are a basic feature offered by most sites. Some sites provide solutions which enhance visualization using selection of font, font size, content and colour. A small number of sites provide point-in-time access to legislatoin. Some research sites or approaches explore the presentation of legislation or bills in radically different forms: such as graph visualizations or as topic colour-coded icons.

In this paper we first review examples of such visualizations and highlight various approaches that are available in official and public good sites in selected jurisdictions.[1] We

---

[1]The visualizations selected are largely confined to the common law world, selectively reviewing sites in Australia, the United States, Canada and the United Kingdom. We have not sought to be be comprehensive in a paper of this length and we have focussed on sampling official and public good sites in jurisdictions available in English language which are most familiar to us. Undoubtedly further insights would be drawn from surveys of online visualization in other jurisdictions, including other cultural and linguistic contexts. A fuller review would also include commercial publication of legislation, which

then briefly present our own visualizations that focus on the enhancement of the visualization of definitions in the parallel domain of legal contracts. Such visualizations are readily transferable to the legislative domain.

The welter of approaches available raises the question of how we may evaluate the utility of a particular visualization. On what basis are we able to suggest, for example, that a basic presentation of text is any worse than a site which provides graphical sliders allowing access to point-in-time versions of legislation? While we intuitively expect more 'advanced' visualizations to be preferable, what are our theoretical or empirical grounds for such conclusions? Furthermore we might ask, better or worse for whom? Current visualizations do not necessarily distinguish between lawyers, citizens, law makers, advocates and other users, who have quite distinct needs. Drawing particularly on the fields of information visualization and knowledge visualization, we conclude by presenting a potential theoretical framework for grounding the visualization of legislation, and discuss the evaluation of legislative visualization.

While a variety of definitions exist of what might be meant by visualization, in the context of this paper we primarily mean *the use of graphics, images or symbols (other than words themselves) to enhance the communication of meaning contained in or associated with (legislative) text.* While primarily in the text itself, meaning in text extends beyond the words themselves, for example information such as document structuring or relationships between concepts found in text. In Section 4 we explore this definitional issue further.

# 2  A Survey of Online Visualization of Legislation

## 2.1  Before the Information Age

Before considering the development of law online, it is worth giving some consideration to the presentation of law in its pre-online forms, as practices from this period have been carried forward in online formats. A concern to increase the accessibility of law has been long standing and continues in many respects to be unresolved. One dimension of such accessibility has been the presentation of law.

Presentation of law in earlier practice of the British Parliament consisted of great slabs of discursive text. In the early-nineteenth century, Jeremy Bentham (credited with being one of the writers influencing later reforms) vociferously critiqued the problems of legislative drafting including the failure to use such obvious tools as division of legislative texts into digestible portions and section numbering to aid retrieval. [15], [11, pp 250-251] Practices such as section numbering and the breaking up of text were officially endorsed with the passage of Britain's first Acts Interpretation Act in 1850 and bedded down after establishment in the late nineteenth century of the first parliamentary drafting office standardised a number of reforming practices in legislative presentation including the structuring of Acts in parts and the use of sub-paragraphing. (See for example Figure 2) [11, p 250], [15, 22, 40]

---

particularly addresses the needs of lawyers and users in large organisations.

In the early nineteenth century law was inaccessible in other senses. A measure of how far things have improved is gleaned from Bentham's observation that "the present price of the last edition of the statutes (Statutes of the Realm) exceeds the average annual income of any individual of the labouring classes in England". [11, p 239]
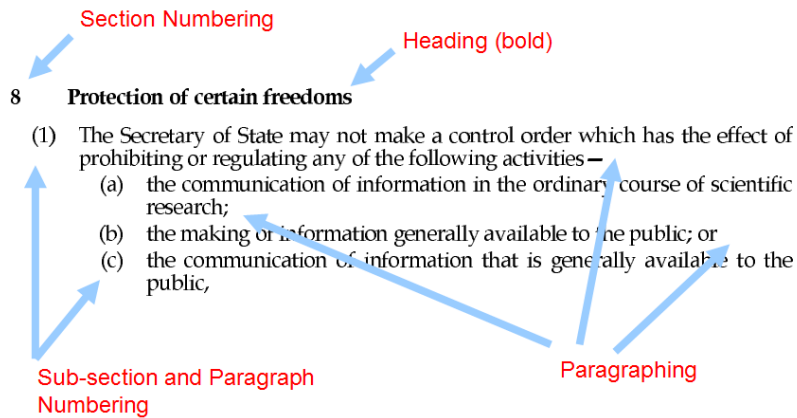


**Figure 2:** Example of now standard presentation of legislation from the UK Export Control Act 2002

Of course we may now obtain access to the Statute book instantly and at negligible cost. The Legal Information Institute (LII) movement has made a considerable contribution to solving this problem. Despite this, legislation (or more precisely its *meaning*) remains in practical terms inaccessible to most citizens of modern democracies for reasons connected with the content and presentation of the law itself.

Bennion, the author of a leading text on statute law makes the following observation:

> "It is strange that free societies should thus arrive at a situation where their members are governed from cradle to grave by texts they cannot comprehend."
> [10, p 8][2]

His observation is not only poetic, it implicitly suggests a standard that legislative communication ought achieve. In 1990, the Law Reform Commission of Victoria in its report titled *Access to the Law - the Structure and Format of Legislation* found it pertinent to quote him. It went on to illustrate the validity of the observation by reporting on reading difficulty measures in application to legislation - demonstrating the inaccessibility of the language to all but a vanishing proportion of the population. Notwithstanding the considerable success in making legislation accessible via online means, and the achievements of the plain English movement in improving the expression of legislation, there is little to suggest that 'Bennion's Conundrum' is substantively less true today than it was in the

---

[2]Ironically, Bennion himself did not believe the problem can be fixed. He complains that he is cited as authority for reform stating 'that it is strange does not mean it is remediable'. Law, he regards as an expert domain and legislation as exclusively addressed to lawyers. The lay person, he says, ought not consider that he or she can understand it any more than the intricacies of medical knowledge without the intermediary of a medical practitioner. [9] Successive New Zealand Commissions (see below) seem to have disagreed with him (particularly the 2008 Commission). It is not an unreasonable expectation that the communications of those who govern us ought be comprehensible. The proponents of plain language convincingly demonstrate the existence of extensive empirical studies establishing the benefits of plain language to comprehension, including in the legal context and the widespread support of plain language measures to increase comprehensibility adopted by legislative drafting offices. [31]

past.

In 1990 the Victorian Commission went on to identify three primary causes of legislative incomprehensibility: defective language; defective organisation; and defective layout and formatting. The second and third 'defects' overlap the field of visualization with which we are concerned. The need to trawl through multiple cross referenced sections or definitions, for example, falls within the topic of organization. Whereas facilities such as headings and numbering impact on how easily law can be found and read.

In respect of layout and formatting (the primary focus of their report) the Commission express concern about typography, density and spacing of material, the indentation of text, the placement of section and sub-section numbers in the body or margins of text and insufficient use of bold font and italics. The size of typeface is another concern. The Commission went on to advance proposals such as the greater use of graphics instead of words where appropriate (e.g. a map or flow chart). They suggested that explanatory material could be boxed and appear together with legislation. They complain that readers need to be 'warned about defined terms' (e.g. by highlighting or special marking). They also suggest the importance of indexes for legislation. They propose the use of decimal numbering systems. They address the incomprehensibility of amending legislation which cannot be read on its own, as it can only be understood if the original legislation is read together with the amending provisions. They suggest instead that the whole amended provision be shown with relevant changes being appropriately highlighted. [1]

Their 'radical' proposals did not go unmarked, though 22 years later they are still largely unadopted in the official presentation of legislation. In 1993, the Australian Commonwealth Parliament inquired into legislative drafting by the Commonwealth, canvassing among other things a number of the issues raised in the *Access to the Law Report*. The Committee appeared to endorse the value of use of graphics. [2, p 153] It expressed support for the use of margin layout for numbering (although calling for empirical assessment of different types of layout). [2, pp 154-155] Proposals for expressing amendments in an integrated form were met with concerns from Office of Parliamentary Counsel as to the additional material that would have to be created. Traditional book publishers supported the system then and still in use, Softlaw Corporation (known for developing the logical representation of legal rules using computational techniques) however supported reform. [2, pp 149-152] Bold font or other marking of definitions similarly received a skeptical response from Parliamentary draftsmen who expressed concern about 'distracting' the readers and legislation being more difficult to read. While attracted to the idea, the Committee wished it to be tested by proper user evaluations. [2, pp 156-157] Decimal numbering systems were rejected by the Committee on the basis of experience from the Queensland Parliament which suggested that parliamentarians found it difficult to use. [2, p 117]

Many of the proposals that seemed overly radical or impracticable in 1993 are in operation in various online sites. The bolding or marking of definitions is for example adopted in AustLII sites. CanLII, in its prototype 'point-in-time' site provides side by side mark up editing showing amendments in context, similar to contextualised amendments proposed by the Victorian Commission. The constraints of paper which earlier made formatting choices problematic are irrelevant to an online environment.

Nonetheless in the formal legal world progress can be slow. As recently as 2008, the New Zealand Law Commission and the New Zealand Parliamentary Counsel's Office together undertook an inquiry into the Presentation of Law starting from the proposition that: *'It is a fundamental precept of any legal system that the law must be accessible to the public.'* [3, p 12] It identified three factors bearing on accessibility: physical or electronic availability, users being able to know where to find the law (navigability) and that the law once found is understandable. [3, pp13-14] The second of these again bears on how law is visualized. The recommendations of this Commission were, however modest, being concerned with issues such as progressing online availability (a problem largely solved or well on the way to solution) and continuing the availability of the law in hard copy form (an anachronism in the 21st century).

## 2.2   Public good publication online

In this section we review the contribution of the LII movement to online publication of legislation.
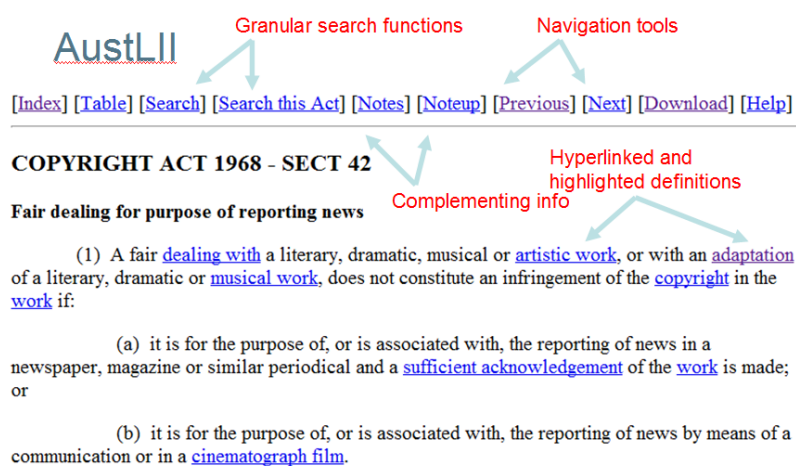


**Figure 3:** Visualization features provided in AustLII legislation

Peter Martin, the co-founder with Tom Bruce of the LII movement observed in 2000 that the most heavily used part of LII was its offering of the U.S. Code, automatically kept up to date by software checking government information sources. In 2000 LII was processing over a million data requests per day: a significant contribution to accessibility of law. LII's origins in 1992 were on a gopher server and the first legislation provided online was the Copyright Act. In March 1993, while snowed in by a blizzard, Martin produced a hand marked up version of the United States Constitution for the World Wide Web. These developments were virtually contemporaneous with the emergence of the World Wide Web itself. At a cost of $250,000 per annum in its initial years, run from Cornell University Law School, LII became a major online provider of open access legal information. [37] It is an example of research driven contribution to society.

These developments were followed quickly by developments in other jurisdictions. By 1 March 1995 AustLII had established a free online site publishing, among other legal resources, Australian Commonwealth statutes and regulations. Also run out of universities (the University of New South Wales and the University of Technology Sydney), its founders, in 1995, stated their philosophy along the following lines:

> *The international development of public legal information servers is part of the more general movement to create publicly available (or 'free to air') resources on the Internet, similar in some respects to the creation of public libraries in the nineteenth century. The Internet is fast becoming home to commercial providers of information, and effective means of charging for even occasional uses of resources are being developed. The countervailing movement, of which AustLII is a part, aims to ensure that some part of cyberspace is public space, where no one is denied use of resources because of financial considerations.* [25]

These early LII's were followed by similar efforts in Canada and Britain (2000) and by 2005 a range of jurisdictions either had their own LII's or were in the process of establishing them. The LII movement also by 2000 had given birth to a global coalition of LII movements known as WorldLII which by 2005 provided free legal information from 55 countries. [30]

The visualization adopted by AustLII (similar to other LII sites) is in part a reflection of its time with an emphasis on hypertext functionality. While relatively uncomplicated, its features appear to be effective, and compare favourably with some of the official sites we have reviewed (some of which are nothing more than text and others which at least at level of personal preference of the authors are not as intuitive and usable). AustLII treats each section as a separate data point and provides key navigational links including: to legislative cross references in the same Act, defined terms, previous and succeeding articles, table of contents, references to materials commenting on a provision, amending notes and Act specific searchability which returns a list of individual provisions in which the search term occurs.[3] (See Figure 3)

AustLII's basic features were discussed by Chung et al. in 2000 including a discussion of an assessment of AustLII against web accessibility heuristics (of interest in the context of evaluation), and a general avoidance of images (of interest in the context of visualization). Chung noted its minimalist philosophy, including the avoidance of images that do not add meaning. Perhaps this view reflected a time when images were widely and injudiciously deployed on web sites and bandwidth constraints meant adding significant time lags. [19]

As with other sites we have reviewed, the CanLII legislative site provides interesting and different features to its users. Funded by Canadian law societies, CanLII aims to provide free access to law. Originally based at the University of Montreal the development of CanLII is now carried on by Lexum, a spinout from the university. It provides a point in time comparison feature that allows the selection of any point in time version and comparison with any other version in side by side layout. Each section of legislation is also hyperlinked to a pop-up window referencing citing cases.[4]

---

[3]For example see AustLII, Commonwealth Consolidated Legislation, http://www.austlii.edu.au/au/legis/cth/consol_act/

[4]CanLII legislation site, http://www.canlii.org/en/info/legislation.html

## 2.3   Official sites

Official publicly accessible sites began to emerge about the same time or shortly after the public good sites discussed above. Greenleaf notes the scaleplus offering of the Australian Attorney-General's department which was available at the time that AustLII was coming into being. [25] Martin notes that in 1995 only ten U.S. states had their legislation available online. [37] Official sites vary widely in their quality and style, although as far as we are aware no user evaluations have been carried out of any particular online framework. This review samples legislation sites in the Australia, the United Kingdom and the United States. The following selective descriptions provide a sense of where official practice lies in the online presentation of law.

### 2.3.1   Selected Australian Jurisdictions

In Australia we compare Federal with selected State and Territory sites each of which adopt different approaches.

The Commonwealth framework presents legislation in a variety of formats, with the principal presentation being through windows giving contemporaneous access to the table of contents as a clickable tree on the left of the page and window containing a html version of legislation on the right of the page. Users can select an alternative visualization providing downloadable versions in pdf, rtf or zip formats, which are represented by document icons. The official and current version of a law is indicated visually by an icon of Australia with a green tick (for example over the pdf version). The presentation is as far as possible identical for each format. The site remains close to the hard copy versions on which it is based. Each piece of legislation is provided as a whole document, rather than being disaggregated into its component sections. The site uses anchors for navigation on its html home page for each Act. An examination of the page source for the webpage indicates that considerable metadata is stored with the legislation, although this functionality is not explicit to the end user. For example a class is assigned to each block of text including classes such as *definition, subsection, paragraph, section and other headings*. This metadata is potentially valuable for a variety of applications. The site is maintained by the Commonwealth Office of Parliamentary Counsel, and is the only **authoritative** site for Australian Federal legislation. The site makes available its content on a creative commons basis and encourages linking and data mining of the text, although access is through the provided versions, rather than via an API for the underlying data.[5]

New South Wales maintains a site closer in form to the AustLII site with the primary presentation being a section by section navigation framework with forward and back buttons. The table of contents appears on the left and can be minimised by dragging. Search functionality is provided. The site also allows the legislation to be presented in a page of different granularity: the entire Act, a part, division or single section. Searchability for a single Act is provided. The page source shows a complex mark up. The site is

---

[5]Australian Office of Parliamentary Counsel, Commonwealth Legislation site, http://www.comlaw.gov.au/

maintained by the NSW Parliamentary Counsel's Office.[6]

The Australian Capital Territory maintains a document oriented site providing access to the current and historical versions of legislation as separate downloadable files in pdf or rtf version. HTML versions are not provided. Great store is placed in the authenticity of the pdf version of the *document* which is digitally signed by the Parliamentary Counsel's Office. Icons representing the various types of document formats available are used to represent these versions on the legislation landing page. The page also provides immediate access to subordinate legislation and accompanying materials such as the explanatory memorandum. The site states its purpose to include making timely and comprehensive legislative information freely available. User friendliness is a stated goal. While a search function is provided, the search returns a whole Act or regulation, rather than a provision in which the search term occurs. (See Figure 4)[7]



**Figure 4:** Document focussed visualization of ACT Human Rights Act 2004

Tasmania's legislation site is again different. It is html based with legislation presented at the level of individual rules and navigation provided through a table of contents on the left of the page, similar to the New South Wales site. Unlike other sites, clicking a division or a part only returns the first section of that portion of the legislation. It is however possible to obtain a html view of the entire Act by clicking the Act title.

The site provides information on how to hyperlink to specific rules in an item of legislation or the legislation as a whole. The site provides point in time access, allowing a user to view a particular rule or the entire Act at a chosen point in time. Also cross-references are activated as hyperlinks. This framework, implemented through a project called Teratext was innovative for its time, being one of the first sites to provide facilities of this kind. The project was sponsored by the Tasmanian government to improve community access to the law, taking legislation from printed (and out of date) consolidations, to online accessible law. Tasmanian law was converted to Standard Generalized Markup Language (SGML) format for this purpose and the electronic database forming the law was recognised by Act of Parliament as the official version of legislation. [6] [5].[8]

---

[6]NSW Parliamentary Counsel's Office, Official NSW Legislation Site, http://www.legislation.nsw.gov.au

[7]ACT Parliamentary Counsel's Office, ACT Legislation Register, http://www.legislation.act.gov.au

[8]Tasmanian Office of Parliamentary Counsel, Tasmanian Legislation Online, http://www.legislation.tas.gov.au

### 2.3.2   Selected U.S. Jurisdictions

The following paragraphs briefly review some of the government maintained sites in the United States.

The United States Code maintained by the Office of Law Revision Counsel of the United States House of Representative has recently[9] been re-released in a beta site which improves its visualization. Colour is used to soften the appearance of the site. Headings are enlarged and presented in browns. The Code can be navigated at title, chapter or article level. Like other official sites, internal hyperlinks are not provided. An examination of the source code suggests presentation is generated by javax programs, presumably drawing from underlying databases.[10] The current version of the US Code online at the time of writing is provided in pure ascii format downloadable as entire titles (i.e. the equivalent of a scroll). (See Figure 5) The site warns readers that the online version of the Code is not official and to refer to the printed version for the 'official' text. Access is provided in pdf format to chronological statutes as issued by the Congress. The older site provides a search function.[11] The beta site is a considerable advance in technology and visualization.



**Figure 5:** Simple ASCII presentation of download of US code

Virginia is another jurisdiction currently undergoing reform of its online presentation of law. A site provides the kind of features seen in the US Code case: use of colour and bold font to enhance readability. Navigation is provided through the availability of previous and next buttons to navigate from section to section. Rare for an official site, the beta site provides internal cross-reference hyperlinks allowing a reader to immediately navigate to a referenced article or to an amending Statute. The table of contents and a search feature are also available on every page. It is not possible to call law at different levels of granularity (e.g. an entire title or chapter). The underlying technology is an '.exe' program which is called by html href tags, presumably interfacing with an underlying database.[12] An examination of the 'classic site' suggests that much of what is described above was already available in the older site, with enhance-

---

[9]As of July 2012

[10]Office of Law Revision Counsel of the US House of Representatives, United States Code Beta Site, http://uscodebeta.house.gov/

[11]Office of Law Revision Counsel of the US House of Representatives, United States Code site, http://uscode.house.gov/

[12]Division of Legislative Automated Systems the Virginia General Assembly, Code of Virginia, http://lis.virginia.gov/000/src.htm

ments largely focussed on adjusting the visual presentation. At article level presentation remains unchanged, except in regard of change of font to increase readability.

Oregon's government provided online site is limited to the 2011 edition of the Oregon Revised Statutes. Readers must themselves refer to the statutes adopted in 2012 to determine the current state of the statute book. Instructions are provided to readers on how to do so. Whole of text searchability is provided for the site but little else. To access the law, the user navigates to the individual chapter of the law which is presented to the reader in its entirety. Examination of the underlying source page indicates that the underlying data is nothing more than a word document that has been converted to html format. This online version is explicitly stated not to constitute the official Oregon Revised Statutes, which are only available in hard copy form by the Oregon Office of Legislative Counsel. [13]

### 2.3.3 The United Kingdom

The United Kingdom's online legislative site is maintained by the UK National Archives and provides access through tables of contents and section level access. Laws can be viewed in html format or downloaded in pdf format. Entire Acts, parts of Acts or single sections can be viewed. At section level forward and back buttons provide ready sequential navigation of provisions. Point in time access is available through a clickable timeline which the user can display.

Geographical application of the law is also visualized through an inline icon which indicates which of the jurisdictions of England, Wales, Scotland and Northern Ireland the legislation applies to. An interesting innovation is the preparedness to use very prominent footnoting and marking to signal changes to the text or relevant commentary. (See Figure 6.)



**Figure 6:** Point in time and edit annotations in UK Human Rights Act visualization

Examination of source pages indicates that underlying data is maintained in an xml schema and css and javascript are used extensively. Information is made freely available under an open government licence (including scripts and data).[14] Unusually (perhaps uniquely for an official site) the UK, through

---

[13]Legislative Counsel Committee of the Oregon Legislative Assembly, Oregon Revised Statutes, http://www.leg.state.or.us/ors/

[14]UK National Archives, the Official Home of the Revised Enacted United Kingdom Statutes,

the Office of Public Sector Information, also provides a legislation API (Application Programming Interface), enabling developers to directly access and republish (or otherwise use) the underlying legislative data. The API pages emphasise that the development of the legislative data base began with an API. It appears to be one of only two official re-conceptualisations of legislation as data rather than law.[15]

## 2.4 Legislative Data and Hacker Visualizations of Law

The work described above has largely been carried out in the context of major institutions, whether primarily universities (in the case of the LII movement), or government agencies in the case of officially maintained sites. The availability of law as data is however a profound change that gives rise to the possibility of groups even as small as one or two people making their own visualizations or analyses of legislative data available. We see some examples of these new possibilities below.



**Figure 7:** Comparison of Presentations of Oregon legislation

Oregon Laws which appears to be maintained by a single individual does not radically reshape the presentation of laws, but does use simple tools to significantly enhance their

---

http://www.legislation.gov.uk/

[15] Office of Public Sector Information of the UK National Archives, Legislation API Developer Site, http://www.opsi.gov.uk/legislation-api/developer/. The second (and earlier) site is that established by the Tasmanian government, which as early as 1996 provided in section 5 of the Act that the Office of Parliamentary Counsel must maintain a database of legislation, and in section 6 that the databased constituted the official version of the law.

presentation. In contrast to the public site, which as we have seen presents the law as 'unofficial' plain text in lengthy chapter by chapter blocks, Oregon Laws provides hierarchical navigation to the section level and uses simple techniques such as heading sizes, colour, bold font and 'more' tags to achieve a visually informative and aesthetic presentation of laws. Frames are used to provide information on related statutory provisions (extracted via citation analysis) and the broader context of a section is contextualised by a block style menu list. Referenced provisions used in an article are hyperlinked and highlighted.[16] (See Figure 7.)

The Virginia Code for Humans site is another small scale 'hacker' visualization of law supported by grants from philanthropic foundations. (See Figure 8.)

It enhances the presentation of legislation by use of features such as colour and font size. It also provides 'cloud' navigation of topics in the legislation and hyperlinking of definitions (a feature also provided on the official beta site). The site provides pop-up visualization of the defining text of defined terms where it is used in a legal rule. This is an obvious enhancement that might be made to the visualiza-



**Figure 8:** Waldo Jaquith's Virginia Decoded with definition text pop-up

tion of law (but does not seem to have been used anywhere apart from this site).[17] The site adopts an explicitly 'open access' philosophy, releasing both underlying data and making the underlying source code for the visualizations developed in php and mysql are released as open source under a GPL licence.[18]

The developer, Waldo Jaquith explicitly states his purpose to be the enhancement of the presentation of legislation 'to make it all more understandable to normal humans'.

> *State codes are wretched. Seriously, look at a few: California's, New York's, Illinois', and Texas' are all good examples of how stunningly difficult that it is to understand state laws. They don't have APIs. Virtually none have bulk downloads. You're stuck with their crude offerings.*[19]

Jaquith's views are hard to argue with. Though there are some developments towards improving government provided online legislative sites, we have seen in our brief survey that they may consist of nothing more sophisticated than an online text dump, with a

---

[16] 2011 Oregon Revised Statutes, https://www.oregonlaws.org/oregon_revised_statutes

[17] Waldo Jaquith, Virginia Decoded - The Code of Virginia for Humans, http://vacode.org

[18] see also Waldo Jaquith, The State Decoded Source Code Repository, https://github.com/waldoj/statedecoded

[19] Waldo Jaquith, the State Decoded About page, http://www.statedecoded.com/about/

linking table of contents and may not even constitute the official version of the law.

## 2.5    Novel and Purpose Specific Visualizations

So far we have examined sites which have largely been motivated by the primary purpose of making laws available. Even for this single purpose, the diversity of approaches is significant. The following visualizations illustrate the potential for legislative data to be visualized in novel ways for a variety of purposes, including direct expression of the meaning of a legal rule.
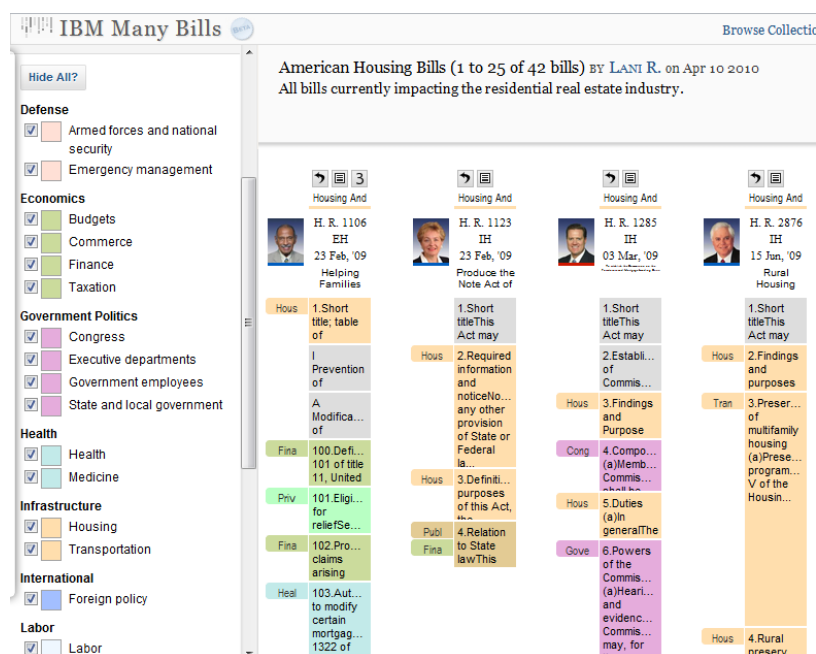


**Figure 9:** IBM ManyBills Topic Icon Visualization

Word clouds are rarely used in formal legislative sites. Such clouds can however be readily found on the internet and employed for the casual review of specific laws.[20]. Critics of the Northern Territory intervention in Australia, which was purported to be for the purpose of addressing child abuse in indigenous populations, point out that the words 'child' and 'children' do not appear in the legislation at all and instead that words related to land and its control appear frequently in the Northern Territory Emergency Response Act. They use a word cloud to illustrate their point.[21]  Matt Stiles at the Texas Tribune uses a word cloud to assess the legislative priorities of legislators, finding most Bills are concerned with community issues such as education.[22]  IBM researchers present a word cloud of the most frequently used words in EU legislation titles.[23]  These presentations seek to make sense of complex legislative data for political or descriptive purposes. The Virginia Code for Humans on the other hand makes traditional use of a word cloud - i.e. a hyperlinked summary of key concepts explored by an item of legislation.

---

[20]By searching for 'legislation' and 'word cloud'

[21]Jens    Korff,    Creative    Spirits,    Northern    Territory    Emergency    Response, http://www.creativespirits.info/aboriginalculture/politics/northern-territory-intervention.html

[22]Matt Stiles, the Texas Tribune, 6 January 2011, Word Cloud Shows Lege Priorities So Far,    http://www.texastribune.org/texas-legislature/texas-legislature/word-cloud-shows-lege-priorities-so-far/

[23]IBM Research and IBM Cognos Software Group, Visualizations: Words Used in Titles of EU Legislation 2011, http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/words-used-in-the-titles-of-eu-leg

ManyBills seeks to address the 'length', 'complexity' and 'obscurity' of bills, providing an interactive environment in which bills and their topic are presented in colour coded 'blocks' which enable the user to distinguish topic areas covered by bills. These topics may be disparate as a result of the political process. The system also provides information on congressional sponsors of bills. It deploys visualization and data mining techniques to extract and present topic related information. Users are able to view Bills at a document and section



**100**   **No entry signs**

A driver must not drive past a no entry sign.

Maximum penalty—20 penalty units.

*Example—*

**No entry sign**

**Figure 10:** Explaining a no entry sign

level. This visualization provides access to law by focussing on graphical presentation and use of colour rather than text to convey information. [7] (See Figure 9.)



(2)  If the driver is to leave the roundabout less than halfway around it, the driver must enter the roundabout—

(a)   from the left marked lane; or

(b)   if the road is not a multi-lane road—as near as practicable to the left side of the road.

*Example—*
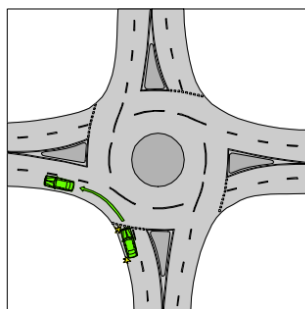
**Example 1**
**Leaving a roundabout less than halfway around it**

**Figure 11:** What lane to use turning left on a roundabout

Bommarito and Katz' work on the United States Code, rather than being concerned with visualization for end user purposes seeks to visualize the United States Code for the purposes of research insight. They particularly seek to assess the complexity of law and the evolution of that complexity over time. They represent the Code as dual overlaid hierarchical and citation networks, exploring a number of properties of these networks, and particularly examining the issue of complexity. [13] [14]

Rasmussen is a small scale commercial venture which describes itself as 'opening up political data, structuring, analyzing and visualization it'.[24] Among its visualization is a flash animation of the development of fields of EU legislation, using growing bubbles to represent the growth of the number of laws adopted in such fields. The purpose of such a visualization is to illustrate their respective importance and growth over time.[25] Rasmussen also take the trouble to provide an API for European Union legislation which interfaces with official EU legislative sources. They describe their API as helping 'you conduct research, create data visualizations or you can even build applications upon it'.[26]

It will be noted by the reader that none of the visualizations examined so far seek to directly present the meaning of legal rules in graphical form.

The work of Haapio and Passera is a demonstration of the potential for visualization to

---

[24]Buhl & Rasmussen, Data Analysis, http://www.buhlrasmussen.eu/index.en.html

[25]Buhl & Rasmussen, the Evolution of European Legislation, http://epdb.eu/eulegislation/

[26]Buhl & Rasmussen, API for European Union Legislation, http://api.epdb.eu/

play an integral role in the expression of legal rules. Combining legal and graphic design skills they illustrate the clarity that is obtained by using graphical features such as a timeline to represent when the contract is in effect and possible termination events, the use of transportation icons to represent passage of risk and ownership of goods in transit and the use of bar charts to convey the relationship between delay and liquidated damages. [27,28,38] Their work is outside the computational context but conceivably computational tools could be created to facilitate the drafting of 'visual rules', or automated translation from text to visual representation, in appropriate cases.

In Australia a legislative case which uses graphics to directly communicate the meaning of legal rules are the Queensland Transport Operations (Road Use Management—Road Rules) Regulation 2009 which are liberally illustrated with explanatory diagrams to enhance the communication of meaning. (See Figures 10 and 11.)

## 3   Visualizing Definition Networks

In this section we briefly outline work undertaken to develop prototypes for the representations of definitions and their associated networks within the context of contracts, which are a parallel domain for legislation. This work will be fully reported in a separate article which has been submitted for publication.[27] It is part of broader research exploring the development of software based tools for enhancing the reading and writing of law. Visualization is a subsidiary portion of the scope of envisaged work, but nonetheless an important one. Previous papers have explored the characterisation of legal language in contracts, through profiling of a corpus of contract documents, [21] and the use of rule based methods, machine learning and hybrid techniques for the automatic classification of text in contracts. [20]

There are of course many potential visualization enhancements that might be addressed. In a contractual context starting with definitions reflects the significant proportion they form of contracts and their role in controlling contract meaning.[28] They also of course play an important role in legislation and directly impinge on the readability and expression of law. Definitions are typically defined in a glossary or definitions section and are then used throughout a legal document. Many sites do not provide any special enhancements to assist in the use of definitions. Some sites provide hyperlinks allowing immediate navigation to a defined term from a legal rule where it is used. The only site however which we are aware of which provides *in context* pop-up access to the meaning of a defined term is the Virginia Code for Humans developed by Waldo Jaquith, which we discuss above.

We have developed a number of prototype visualizations of definitions and their associated networks that may be of assistance to readers or drafters of contracts or legislation. We describe them briefly.[29]

---

[27]Michael Curtotti, Eric McCreath, Srinivas Sridharan *Software Tools for the Visualization of Definition Networks in Legal Contracts* - under peer review.

[28]One of the author's experience in the negotiation of contracts as an in-house lawyer was the genesis of the identification of definition networks as an area where visualization may be useful.

[29]These visualizations can be viewed at http://buttle.anu.edu.au/contracts/

A. **Definition Network Graph Visualization:** Definition networks sometimes form networks where one definition will refer to another in how it is defined. Such graphs can be complex (even in contracts). This visualization provides a node and link diagram which expresses such connectivity between defined terms. This kind of visualization is potentially useful to readers in seeing the relationships between defined terms used in a legal document. In contrast to the multi-layer pop-up (described below) it provides an immediate single view access to an entire definition network.

B. **Single and Multi-Layer Pop-up Defined Term Navigation:** As we saw above the Virginia Code for Humans is the only site of which we are aware that uses pop-ups to allow the meaning of a defined term to be accessed in context. We provide similar functionality, but also provide a prototype visualization allowing direct multilayer navigation through a definition network from the rule where the defined terms are used. This is potentially of particular use to readers who often have to navigate a complex network of referenced material before the meaning of a rule can be understood. After navigating three or four pages away from the rule (either through hyperlinks or by scrolling through a text document) the reader is seeking to comprehend, comprehension can be considerably reduced.

C. **Definition Clouds:** Use of word clouds in respect of legislation is rare as we have noted above. There is however considerable scholarship evaluating the utility and effectiveness of word clouds. [8, 29, 36] Features such as placement, font, colour and size of text of varying effectiveness. We develop a 'definition cloud': which although using the same basic arrangements, is conceptually distinct in that it leverages from an existing human created ontology within a legal document. We use size to indicate frequency of usage and traffic light colour scheme to indicate how much of the meaning of a definition is hidden in defined terms used in the definition. A comparison with a more traditional word cloud is provided.

D. **Usage and Obfuscation Icons:** An alternative representation of the same information is provided by a small circular icon placed on the left of each definition. A pie chart in this case indicates how much of the text is 'hidden'. A number placed over the icon provides a usage metric, indicating importance of the provision. This and the previous visualization are potentially useful to drafters in reducing the complexity of their drafts and to readers in identifying defined terms that require particular attention.

E. **Definition-Rule Network Matrix Visualization:** This visualization draws on a tool used in social network analysis and provides a summary representation of the relationship between defined terms and the rules in which they are used. These relationships can signal semantic and topic connectivity between potentially separate parts of a legal document. Weight of relationship is indicated by the darkness of each matrix square.

Such visualizations are largely novel in application to legal documents but their development has been guided by experience in working with legal documents and are for that reason appealing (at least to their creators). This is by no means a sufficient basis for expressions of confidence in their value and below we discuss the topic of evaluation of visualizations, and one of our goals is to carry out such evaluation in future research. Future work will also include extension of such tools to a demonstration legislation site.

# 4 Towards a Theory of Legislative Visualization?

As evident from the review above, practice in the visualization of legislation is diverse. There does not seem to be a theoretical framework grounding such visualization. Indeed, there does not appear to be a shared consensus, or even a developing discourse, on what 'good practice guidelines' might look like. Is it really the case, for example, that the dumping of ascii text online in a bulk presentation of law is not as good a visualization as a presentation which presents formatted individual rules? The absence of such theoretical frameworks means that is difficult to pursue the systematic enhancement of online visualization of legislation: standards and approaches appear arbitrary as they are not non-verified. For these reasons a theory or at least empirical results as to a body of 'good practice' is required to enhance visualization.

## 4.1 Visualization Theory

The absence of a theoretical framework for the visualization of legislation is related to the broader problem in information visualization as a whole. Researchers in the information visualization field describe the absence of theory, or agreed theory. As information visualization ranges over disparate fields the problems are compounded by the diversity of disciplinary constructs that might be applied. Is information visualization to be approached as a problem of cognitive science taking into account the way the brain processes visual information? [4], [43] Or as a problem in information theory emphasising the transmission and encoding of information? [18] Perhaps decision science, which asks how individuals make decisions, applying concepts of utility, should be applied to design questions? [45] Further we have choices such as approaching the problem as one of 'science' with a focus on discovering the principles of information visualization or one of 'engineering' with a focus on application of principles to the creation of software artefacts. [24]

Underlying such questions are even more basic questions of what we mean by visualization. Particularly what do we mean by visualization in the context of legislative documents? We may note that text (consistent with its origins in speech) is a sequence of symbols (words) used to convey meaning. It is essentially one dimensional and meaning is conveyed to the reader as symbols are sequentially parsed. Although a reader 'sees' the text, the processes by which the reader absorbs information from text are adapted from the how we process speech. Visualization, by contrast, is grounded in sight rather than hearing/speech. Ordering may be important or irrelevent in the visualization context. A visualization may be one, two, three or four dimensional (if changing over time). How we extract meaning from what we see, is significantly different from how we extract meaning from text.

In this broad but general context, definitions associated with the concept of visualization are better understood. Burkhard defines knowledge visualization in the following terms:

> *Knowledge visualization examines the use of visual representations to improve the transfer of knowledge between at least two persons or group of persons.* [16]

Legler and Eppler suggest the following definition for a 'visualization method':

> *A visualization method is a systematic, rule-based, external, permanent, and graphic representation that depicts information in a way that is conducive to acquiring insights, developing an elaborate understanding, or communicating experiences.* [34]

Information visualization is defined as the computer assisted use of visual processing to gain understanding. [16, 17, 34]

Tegarden notes that visualization 'allows decision-makers to use their natural spatial/visual abilities' and that it 'exploits the human visual system to extract information from data'. [44]

For the purposes of this paper, drawing on ideas such as these by 'visualization' we intend the use of graphics, images or symbols (other than the words themselves) to enhance the communication of meaning contained in or associated with (legislative) text.

## 4.2   Suggestions for Framing Legislative Visualization

Given the diversity of approaches potentially available it is reasonable to focus on those ideas that intuition suggests may be particularly useful or relevant to the task of visualization legislation. Of course this statement carries assumptions – it is an applied approach – prioritizing specified task oriented outcomes over a general theoretical framework. For our purposes (enhancing the visualization of legislation) this makes sense. Grinstein presents a conjecture for addressing visualization problem of this kind. [26]

> *Given a data set $D$, given a task $T$, for a given display, there exists a visualization $V$ such that the perceived information $I$ is such that task $T$ is optimally perceptually/cognitively "resolved". This means that no other visualization will solve task $T$ as well or that the perceived information $I$ is the best for resolving task $T$.*

While such mathematical formalism will not necessarily appeal to all tastes, it is a useful crystallisation of the issues that need to be addressed in developing a visualization and helps us clearly express our intended meaning. As Grinstein notes a number of terms are undefined or unknown or measures to be defined.

While generally useful for the visualization of legislation we can adapt it to be more precisely targetted for our needs:

> *Given a set of legislation $L$, given a user $U$ with task $T$ requiring knowledge of $L$, there exists a visualization $V_1$ with features $F_1$ to $F_n$ such that the perceived knowledge $K_p$ conveyed to $U$ for task $T$ is better perceptually/cognitively "resolved" than in a given set of visualizations $V_2$ to $V_n$.*

If our interest is the performance of a particular task or achieving a particular outcome, such as increasing a user's access to law, then such a framework suggests how we might practically and systematically pursue such a desired outcome to enhance visualization

19

outcomes. It draws us to questions such as the following: Who are our intended users? What are their needs and relevant characteristics? What tasks would they intend to carry out. What visualization features are to be provided? For enhancing the communication of what knowledge? How effective are they such communication as compared with existing visualizations?

The inclusion of the user U in the framework allows us to distinguish between different users of legislation - such as drafters, parliamentarians, legal and professional advisers, activists, researchers or members of the public, whose tasks T, will differ and will be conditioned by their intended outcomes, pre-existing knowledge and needs. We see, for example, that the visualizations produced by Bommarito and Katz, who were seeking to visualize the complexity of an entire body of legislation, address an entirely different task to that which might face a legal adviser who may for example be seeking to predict legal outcomes against given facts.

Confining the problem to comparison of a set of given visualizations enables us to undertake empirical evaluation against existing legislative visualizations (such as those surveyed above) as opposed to developing a theoretically 'optimal' model. How such optimality might be discovered in the abstract is not necessarily evident.

The substitution of a knowledge parameter for the information parameter, emphasizes a 'knowledge visualization' framework which again is more suited to our needs. Although Grinstein notes that the end point in information visualization is 'does the user get it?', and the two concepts are closely related, generally information visualization is focussed on the extraction of new insights from data, whereas the explicit purpose of knowledge visualizations as conceptualised by Burkhard is to improve the transfer of knowledge between individuals or groups. [16] This focus is again congruent with the legislative context, where for example government may be seeking to make knowledge available to the citizen, either to ensure the citizen knows their rights, or complies with the law.

The inclusion of features (F) in our model again enables us to focus on those elements that contribute to performance for a given task. For example point-in-time access may be particularly valuable to a lawyer seeking to litigate a case based on events that occurred some years ago, it may be less pertinent to a member of the public seeking to know the current state of the law.

## 4.3   Evaluating Legislative Visualizations

Having such a framework we may also observe that there is a universal baseline against which any online visualization of the law may be tested - i.e. the hard copy version of the legislation. At a minimum we would wish task T to be better resolved in an online visualization than for a user using the paper text of a law.

Of course central to such a model are empirical studies of visualizations, which are widely agreed to be essential to progressing information visualization in a coherent fashion. [24, 26, 35, 39, 45]

The need for empirical evaluation is not a new insight. As we have seen, the Australian

Parliamentary Committee charged with reviewing the drafting of legislation in 1993 realized that user evaluations were critical. They considered that human testing would be preferable to any automated testing against metrics, and urged that testing be given priority. They, in particular, rejected readability metrics as of assistance in this context, noting that such measures correlated poorly to real world issues with readability of legislative provisions. In evidence presented to the Committee, the Commonwealth Office of Parliamentary Counsel stated that they planned to undertake a program of testing of a limited number of statutes for the purposes of comparing plain language and then current drafting styles. [2, pp 98-103] Krongold discusses the importance of testing whether individuals are able to understand legislation, preferably during drafting. Like other observers, Krongold is critical of the value of tests such as the Fleisch scale or FOG index, noting that such measures are neither adequate or accurate for legislative documents. She particularly notes that although a negative score may be accurate, a positive score using such indexes has little relationship with actual readability. She notes further that the metrics were developed for general prose and were never tested on legislation. [32, pp 544-545]

Reported evaluations of presentational or visualization aspects of legislative text are very rare. A study of this kind is reported by Stewart who is particularly concerned with section headings in legislation and undertakes usability testing to test two propositions: that the redrafting of section headings as questions would assist in comprehension, and that the addition of headings for subsections would assist in comprehension. Metrics used in undertaking the tests included time on task and accuracy of responses. [42] Another study which combines presentational improvements with plain language enhancement of the Canadian Employment Insurance Act finds clear usability benefits from the combination of such features. Graphical features included use of colour, bold font, font sizing and careful font choice and underlining of defined terms. [12, 23] Passera notes that the combination of plain language enhancements with these visualization changes means that it is not possible to determine the specific contribution made by visualization enhancements as they were not measured separately. Passera's own evaluation of the effect on usability of visualizations in the parallel domain of contract visualization establishes the usability improvements associated with use of diagrams and charts, improving typography and layout out, highlighting key terms, utilizing color-coding in a redesigned table of contents. She tests and confirms four hypotheses: that visualizations support faster reading and more accurate understanding, provide a more positive experience than text only contracts and affect user expectations positively. [38]

A notable officially undertaken evaluation of visualization together with plain language features was a 2010 survey commissioned by the Australian Commonwealth Office of Parliamentary Counsel (OPC) as to user evaluations of enhancements to presentational features of Australian law. The First Parliamentary Counsel notes that 'during the 1990's, OPC experimented with a range of innovations to our drafting style'. The evaluation is limited to professional user groups (judges, tribunal members and their associates, lawyers, Commonwealth employees involved in instructing or advising on law and parliamentary officers).[30] The selection of 'users' is significant and reminds us of Bennion's words as to the audience for whom laws are (implicitly or explicitly) written. Ideally

---

[30]First Parliamentary Counsel Letter *Survey of Users of Legislation 14 December 2010 http://www.opc.gov.au/plain/pdf/2010LegislationSurveyLetter.pdf*

(given the communication and open access considerations discussed above) such surveys would include members of the public with varying experience of use of legislative materials. An indication that a broader survey would provide additional insights is indicated by the divergence in response between members of the legal fraternity as opposed to Commonwealth employees (the latter more positive on average than the legal group in respect of all innovations). The essentially 'offline' conception of the evaluation is notable. None of the questions asked as part of the survey, mentioned or addressed issues specific to the online provision of law (e.g. users experience of using the online site), despite the transition to the online environment being the most significant delveopment of the period under review.

Nonetheless the survey stands out as one of the few official systematic evaluations and validation of selected innovative visualization (and other) elements in legislation. The survey enabled the OPC to validate certain innovations providing evidence for continuing and extending their use and called into question others, against the needs of a selected user group. Features such as the new format for legislation; the new form of commencement provisions; the use of notes; the use of tagging of concepts; the use of tables; and the use of subsection headings rated well. The survey also provided an opportunity for OPC to seek the opinion of users as to other potential enhancements that could be made. Among innovations suggested for consideration which are relevant to this paper were: use of hyperlinks for online versions of laws and 'an online layout compatible with word' (e.g. the need for ready reproducibility in legal advice). Also private sector legislation users mentioned that 'what they needed was clear and plain legislation ... legislation had become overly complex'. Interestingly the user group did not respond positively to 'diagrams', the use of which accordingly is to be reduced. [41] Passera's results are perhaps an indication that the needs of business users may differ from those professionally immersed in the use of legal texts.

In the information visualization field, there are well developed methodologies for evaluation with a variety of approaches available. Plaisant et al discuss a number of them including controlled experiments comparing design elements, usability evaluation of a given tool to refine its design, controlled comparative evaluations of two or more tools (a common type of study) and case studies of tools in realistic scenarios. [39] These kinds of studies are common to research methods in human computer interaction which include also methodologies such as diaries (as a tool for evaluating user experience), focus groups and interviews, automated collection of data, and ethnographic studies. [33] Such literature describes the well established and widely utilised methodologies available for evaluation and we have proposed above one possible framework against which evaluation of online legislative visualizations might be carried out, although undoubtedly others can be imagined.

The apparent dearth of evaluations of the online visualization of law is curious. One reason for the limited number of such studies may be unsurprising - cost and difficulty - a factor noted in the Commonwealth Parliamentary inquiry. Another may be the lack of coherent theoretical frameworks described above. It may however simply be a function of the cultural patterns which develop (or not) within a particular community of practice. However given the well established practice of evaluation within the field of human computer interaction, and a similarly well established tradition of evaluation that the plain language movement brought to the offline improvement of legislative texts, [31]

greater attention is warranted as to how evaluation might be systematized as standard practice in online publication of legislation.

# 5 Conclusions

This paper has selectively reviewed online visualization of legislation in the context of pursuit of open access to the law. This has highlighted the diversity of practice in the field and the absence of measures or standards for determinging the quality of online legislative visualization. The review also indicated that some official sites lag behind the state of the art (although the state of the art itself is unevaluated). The Law Via the Internet movement remains among the most prolific providers of online legislation through its LII sites, despite the gradual progress evident in work to enhance official sites. Some sites provide examples of the potential for easy enhancements that might be adopted by other sites or for radical new visualizations of law to improve access for existing users, address new user needs or enable new questions to be addressed. The essentially arbitrary nature of the various approaches illustrated and the wide range in quality enables us to conclude that there exists significant scope to enhance the online visualization of legislation.

Of particular interest is the emergence of the first official legislative data repositories and associated API's which are conceived as such. This may prove in the long run to be among the more significant developments of recent years, as it provides the potential for legislation to be accessed or analyzed in ways only limited by human imagination. Sites such as Oregon Laws and the Virginia Code for Humans, illustrate that even small groups of actors with limited resources can provide original and enhanced visualizations once the data itself is accessible. Considerations such as these suggest that it would be useful for official providers of legislation to be encouraged to provide and maintain access to legislative data sets together with API's to assist developers (and researchers) to access that data for a wide variety of purposes.

We report prototype visualizations of definition networks in the parallel domain of contracts and discuss the potential for these visualizations to provide useful tools for users of legal documents. The visualization should only be taken as an example of one dimension in which legislative visualization is open to enhancement. Assessing the merits of such visualizations and empirically identifying the range of potential enhancements that better meet user needs can only be reliably determined on the basis sound theory and empirical evaluation as we have argued in this paper.

The exploration of a potential theoretical framework and the evaluation of visualizations addresses the need to systematise approaches and learning that could be drawn from the current diversity of online legislative visualizations. Without a systematic approach to determining what is 'good' visualization, development will likely be haphazard and progressive enhancement will proceed more slowly than might otherwise be the case. Developing a coherent body of knowledge in respect of visualization of legislation and its evaluation may contribute significantly to further advancing accessibility for users of online legislative sites.

There has been remarkable progress in the last 20 years in furthering accessibility to law

and the contrast with past exclusion of the general population from the law could not be greater. Making law available online for free is an astonishing achievement by the people who have brought it into being. The work of the plain language movement has also been a significant feature of the last years. Yet Bennion's Conundrum remains: *'we find ourselves in the profoundly undemocratic situation that we are governed from cradle to grave by laws that most of us cannot comprehend'*. The platforms established in the last twenty years, apart from their direct benefits, offer a departure point for addressing (or further addressing) other dimensions that impede accessibility to law (such as how law is expressed, how it is organised, how it is visualized). It is clear that there are many other possibilities that might be explored, including tools to transform what has largely been a static and formalised process of unidirectional communication of law from government to the governed, into an ongoing conversation between law makers and law users (and among law users themselves) as to the legal rules under which we live. It would not be difficult, for example, for tools to be provided to allow legal rules to be rated by online users, or for comment facilities to provide direct and immediate feedback to those who write the law, and valuable information for parliamentarians.

In future work we intend extending our current work on contracts to legislation and undertaking evaluation of the benefits or otherwise of proposed enhancements or tools, such as tools for visualization of definition networks.

# References

[1] Access to the Law - the structure and format of legislation. Technical Report 33, Law Reform Commission of Victoria, 1990.

[2] Clearer Commonwealth Law: Report of the Inquiry into Legislative Drafting by the Commonwealth. Technical report, House of Representatives Standing Committee on Legal and Constitutional Affairs, 1993.

[3] Presentation of New Zealand Statute Law. Technical Report 104, New Zealand Law Reform Commission and New Zealand Parliamentary Counsel's Office, 2008.

[4] E. Anderson, C. Scheidegger, and C. Silva. The Role of Perception and Cognition Costs in Models of Visualization Effectiveness. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

[5] T. Arnold-Moore. Point in time publication for legislation (xml and legislation). In *Proceedings ot the 6th Conference on Computerisation of Law via the Internet, Paris, France (December 2004)*, 2004.

[6] T. Arnold-Moore, J. Clemes, and M. Tadd. Connected to the law: Tasmanian legislation using enact. *Journal of Information Law and Technology*, 1, 2000.

[7] Y. Assogba, I. Ros, J. DiMicco, and M. McKeon. Many bills: engaging citizens through visualizations of congressional legislation. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 433–442. ACM, 2011.

[8] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202. ACM, 2008.

[9] F. Bennion. If it's not broke don't fix it: A review of the new zealand law commission's proposals on the format of legislation. *Statute L. Rev.*, 15:164, 1994.

[10] F.A.R. Bennion. *Statute law.* Oyez, 1983.

[11] J. Bentham. Nomography, or the art of inditing laws. *The Works of Jeremy Bentham (ed. J. Bowring)*, 3:231 et seq, 1843.

[12] D. Berman. Toward a New Format for Canadian Legislation: Using Graphic Design Principles and Methods to Improve Public Access to Law. `http://www.davidberman.com/NewFormatForCanadianLegislation.pdf`, 2000.

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

[14] M. J. Bommarito and D.M. Katz. A mathematical approach to the study of the united states code. *Physica A: Statistical Mechanics and its Applications*, 389(19):4195–4200, 2010.

[15] F. Bowers. Victorian reforms in legislative drafting. *Tijdschrift voor Rechtsgeschiedenis*, 48:329, 1980.

[16] R.A. Burkhard. Learning from architects: the difference between knowledge visualization and information visualization. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 519–524. IEEE, 2004.

[17] S.K. Card and J. Mackinlay. The structure of the information visualization design space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 92–99. IEEE, 1997.

[18] M. Chen and H. Jaenicke. An information-theoretic framework for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1206–1215, 2010.

[19] P. Chung et al. A defence of plain html for law: Austlii's approach to standards', 2000 (1). *The Journal of Information, Law and Technology (JILT)*, pages 00–1.

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

[21] M. Curtotti and E. McCreath. A corpus of australian contract language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law 2011*, 2011.

[22] M. Evans and R.I. Jack. *Sources of English Legal and Constitutional History.* Butterworths, 1984.

[23] GLPi and V. Smolenka. A Report on the Results of Usability Testing Re-

search on Plain Language Draft Sections of the Employment Insurance Act. `http://www.davidberman.com/wp-content/uploads/glpi-english.pdf`, 1964.

[24] L. Grammel and M.A. Storey. Towards a foundation for information visualization engineering. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

[25] G. Greenleaf, A. Mowbray, and G. King. Public legal information via internet: Austlii's fist six months. *Austl. L. Libr.*, 3:144, 1995.

[26] G. Grinstein. Arguments for a theory of visualization. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

[27] H. Haapio. Contract clarity through visualization–preliminary observations and experiments. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 337–342. IEEE, 2011.

[28] H. Haapio and S. Passera. Reducing contract complexity through visualization-a multi-level challenge. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 370–375. IEEE, 2012.

[29] M.J. Halvey and M.T. Keane. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314. ACM, 2007.

[30] Miller J. The development of the legal information institutes around the world. *Can. L. Libr. Rev.*, 30:8, 2005.

[31] J. Kimble. Answering the critics of plain language. *Scribes J. Leg. Writing*, 5:51, 1994.

[32] S. Krongold. Writing laws: Making them easier to understand. *Ottawa L. Rev.*, 24:495, 1992.

[33] J. Lazar, J.H. Feng, and H. Hochheiser. *Research methods in human-computer interaction.* John Wiley & Sons Inc, 2009.

[34] R. Lengler and M.J. Eppler. Towards a periodic table of visualization methods for management. In *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*, 2007.

[35] Z. Liu and J. Stasko. Theories in information visualization: What, why and how. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

[36] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. *Human-Computer Interaction–INTERACT 2009*, pages 392–404, 2009.

[37] P.W. Martin. The mushrooming virtual law library on the net. In *Cornell Law Forum*, volume 27.

[38] S. Passera. Enhancing contract usability and user experience through visualization-an experimental evaluation. In *Information Visualisation (IV), 2012 16th Interna-*

*tional Conference on*, pages 376–382. IEEE, 2012.

[39] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

[40] D Renton. The preparation of legislation - report of a committee appointed by the lord president of council, 1975.

[41] Orima Research. Office of Parliementary Counsel results of 2010 legislation users survey.

[42] G. Stewart. Legislative drafting and the marginal note. *Statute Law Review*, 16:21, 1995.

[43] O. Swienty and M. Takatsuka. From subjectivity to objectivity in information visualization: challenges in bridging the gap between cognitive theory and information design. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

[44] D.P. Tegarden. Business information visualization. *Communications of the AIS*, 1(1es):4, 1999.

[45] J.S. Yi. Toward practical information visualization theory: Lessons from the history of decision science. *VizWeek 2010: The Role of Theory in Information Visualization*, 2010.

# 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.

---

[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

---

[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.

**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 |

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.

### 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

---

[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]).

## Log of Tokens per Document



## Log of Vocab Size of Document



## Tokens per Document
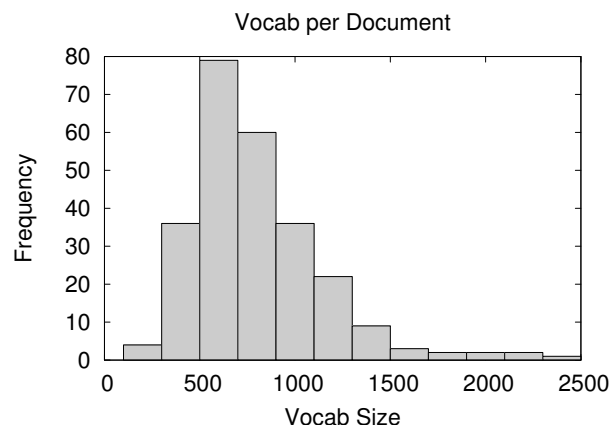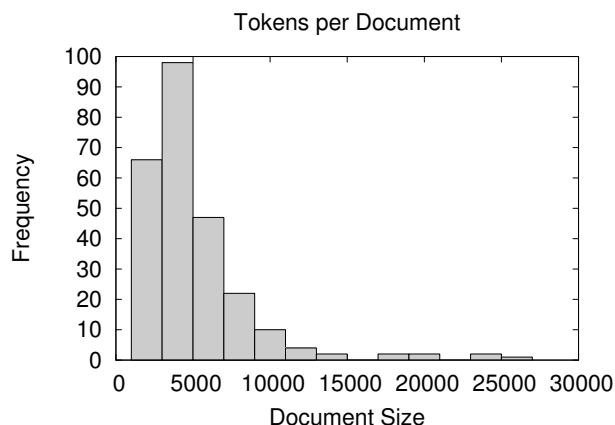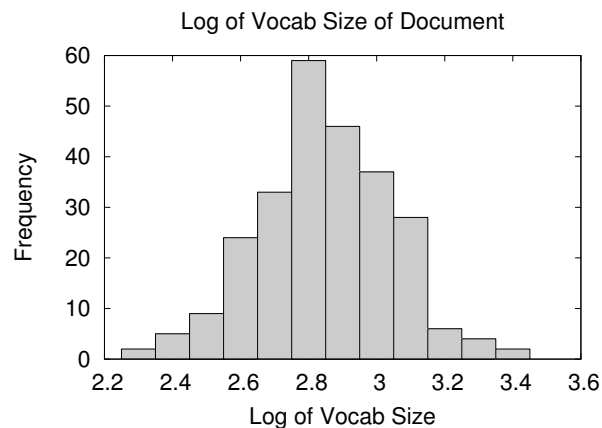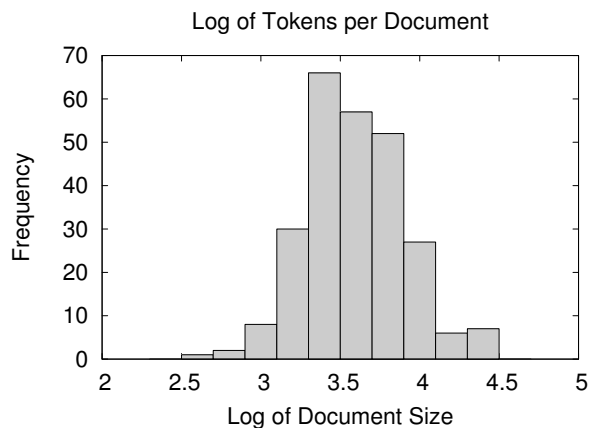


## Vocab per Document



**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.
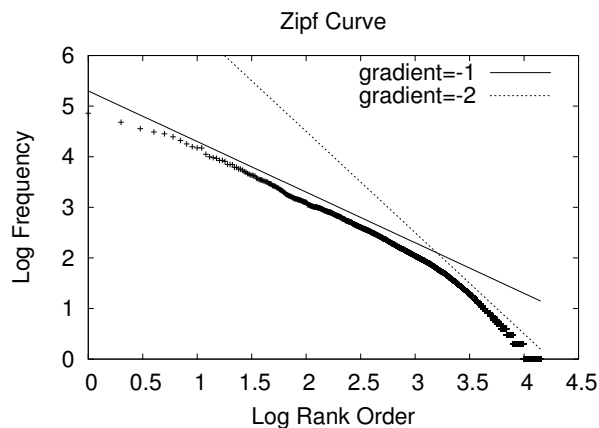
## Zipf Curve



**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

---

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].

of contracts included both function terms and terms that domain knowledge might suggest would be distinctive.

## 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.

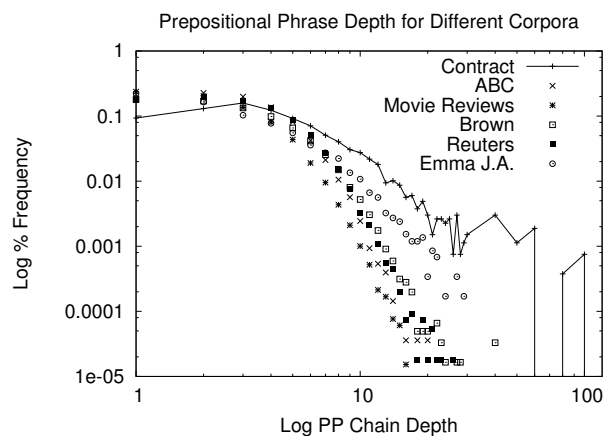[3] L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54, 2009.

[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.